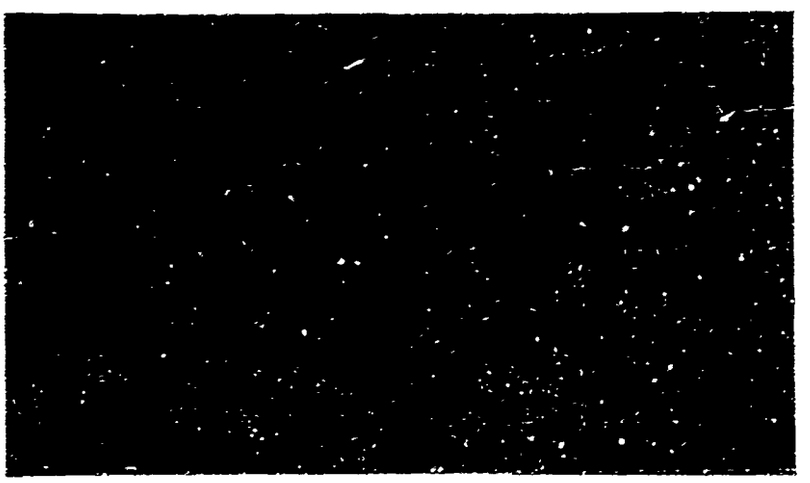
AFOSR 68-2162

AD 677210



1. This document has been approved for public release and sale; its distribution is unlimited.

THE SHUFORD-MASSENGILL CORPORATION

SM

51

AIRMAN QUALIFYING EXAMINATION-66 ADMINISTERED AS A CONFIDENCE TEST

Emir H. Shuford Jr. and H Edward Massengill Jr.

AIRMAN QUALIFYING EXAMINATION-66 ADMINISTERED AS A CONFIDENCE TEST*

Emir H. Shuford, Jr. and H. Edward Massengill, Jr.

Within the past two decades, several developments have occured which suggest that it may be possible to make great improvements in ability, aptitude, and achievement testing. Gains have been made in aptitude and achievement testing, but during the past five or so decades of testing they have resulted from efforts to improve the development and selection of test items for use within the standard framework for objective and semi-objective testing utilizing multiple-choice and completion-type item formats. The new developments, however, call for a fundamental change in the way that a test is administered. They allow for a more penetrating method of measurement which moves ahead of the traditional choice response of the examinee to more directly measure the value and character of the information that would lead to a choice and they measure it in terms of subjective probability of correctiness or, equivalently, degree of confidence.

Although there have been many attempts in the past to measure degree of confidence, to allow partial knowledge, and to eliminate guessing (all of which have met with notable lack of success) there are still reasons to believe that it is possible to improve upon the choice method of test administration.

At first these reasons were logical and mathematical. After the fact it is clear that the first requirement that any method of measuring an examinee's confidence must meet is that it be in his best interest to honestly express his degree of confidence. Toda (1963) in Japan, de Finetti (1961) in Italy, van Naerssen (1961) in Holland, and Roby (1965) in the United States independently discovered special ways of rewarding a subject according to his assignment of confidence to the alternatives in a choice problem. These scoring systems all had the special property that the subject could maximize his expected score *if* and *only if* he honestly expressed his degree of confidence in the alternatives. Shuford, Albert and Massengill (1966) further rationalized these scoring systems and extended their use to the fill-in-the-blank or completion-type item. It is significant that when the previous and unsuccessful attempts to measure degree of confidence are examined as to the mathematical properties of their scoring systems, none of them have been found to use an admissible scoring system. In each case, the scoring system was either so ill-defined that it constituted a projective test to the examinee, or if well-defined, the scoring system had the property that the examinee would make a better test score by not responding with his actual degree of confidence. Although this may not explain completely the reasons for failure of the earlier studies, it certainly gives us reason not to be discouraged by the long run of negative findings with respect to the promise of confidence testing.

Recognition of the new-found promise of confidence testing lead to the occurance of two significant events. The Shuford-Massengill Corporation undertook the development of training and response aids and other materials and procedures which would make confidence testing feasible and suitable for wide spread

*The Fourth Semi-Annual Technical Report (which covers the period November 1967 through April 1968) of work performed under contract number AF 49(638)-1744, ARPA Order Number 833, by The Shuford-Massengill Corporation, One Wallis Court, Lexington, Massachusetts 02173

use in the public school as an improved way of administering teacher-made classroom tests and quizzes. The Advanced Research Projects Agency of the Department of Defense contracted with the Corporation to develop the decision-theoretic psychometrics necessary to guide and support applications of confidence testing to military operations. This contract effort yielded numerous reports (Massengill & Shuford, 1966; 1967; 1968; Shuford & Massengill, 1966a; 1966b; 1967a; 1967b; 1967c) which spelled out the conditions under which confidence testing could yield large gains in personnel selection, classification, training, and education. In response to the promising indications of this theoretical work, the Advanced Research Projects Agency expanded the contract effort to include collecting and analyzing comparative data on the performance of choice and confidence testing with military selection and classification tests. This Semi-Annual Technical Report is devoted to the analysis of the performance of confidence testing in the administration of the multiple-choice items of the Airman Qualifying Examination-66 which is currently used by the Air Force Recruiting Service for the selection and classification of non-prior-service applicants to determine their enlistment qualifications and aptitudes.

PROCEDURE

As a result of discussions with Robert B. Stephens of Hq. USAF and Bart M. Vitola of the Personnel Research Laboratory, we had planned to readminister the multiple-choice portion of AQE-66 to about 300 basic airmen in training at Lackland Air Force Base. Each basic airman devotes half a day to participating in the experimental testing program of the Personnel Research Laboratory. The airman has taken the AQE-66 at a recruiting station or at a high school prior to entering service with the Air Force. By retrieving the original test answer sheets of the airmen, we could compare the data from the original administration of AQE-66 as a choice test with the data obtained from the readministration as a confidence test. This comparison could apply only to the 150 multiple-choice items of the AQE-66. The other test items are speeded computational problems for which confidence testing would not be appropriate. All of the 150 multiple-choice items are five alternative with the exception of several three and four alternative items.

Upon examining the materials for confidence testing, however, the staff of the Personnel Research Laboratory expressed reservations as to the feasibility of the method, since the airmen would represent a broad cross section of educational levels and academic backgrounds. If the airmen failed to clearly understand the rationale and procedures of confidence testing, then the resulting confusion would get in the way of their responding in a meaningful fashion and disrupt the testing process to the extent that the possible gains from confidence testing would not be realized. Realizing that we would have only a three to three and half hour period to train the airmen in how to take a confidence test and to have them complete 150 five alternative items, many of them of considerable complexity and difficulty, we found this arguement somewhat compelling. We proposed, therefore, that instead of risking a larger scale program using 300 airmen we scale down the first testing to include two small groups, each of which would represent a cross section of airmen currently in training at Lackland Air Force Base. This plan was agreed to by the Personnel Research Laboratory and the experimental testing took place January 26, 1968.

Test booklets and testing facilities and personnel were provided by the Personnel Research Laboratory. Mr. C. L. Cannon and Sgt. I. T. Busby, of the Personnel Research Laboratory assisted in the administration of the test. Thirty airmen were tested in the morning and 32 were tested in the afternoon. Edward Massengill of the Corporation spent approximately one hour with each group instructing them how to take a confidence test. The airmen were then given a fifteen minute break and returned to take the 150 multiple-choice items of the AQE-66.

The airmen were allowed one hour and forty-five minutes to take the test. This is the time normally allowed for taking this portion of AQE-66. Some of the airmen completed all 150 items well ahead of the time limit, while others took the full hour and forty-five minutes and completed less than the full set of items. In the original administration of this test, many of the airmen also failed to complete all 150 items. The major difference between the two administrations is that in the original administration, the airmen who failed to complete the test tended to skip many items, whereas, in the readministration, airmen tended to answer all items consecutively up to the point where the time limit was imposed.

All in all, this test was speeded for many of the airmen both in the original administration and even more so when it was administered as a confidence test. The mechanics of taking a confidence test require more manipulation and writing on the part of the examinee than in the case of a choice test. Further, experience indicates that confidence testing leads the examinee to think much more carefully about each test question and its answers than in the case of choice testing. When a confidence test is speeded, there is less time to consider carefully all the relevant information pertaining to a test question and to carefully evaluate this information in terms of how much confidence is justified in each of the possible answers. In addition, the easiest response pattern to develop in a confidence test is that in which all the confidence is placed on one of the answers and no confidence is placed on the remaining answers. These considerations imply that speeding a confidence test will make the data look more like that of a choice test where, in effect, all confidence is placed on one and only one of the possible answers.

Each airman used the *SCoRule*™ response aid to develop his response to the test items and then copy the appropriate letters into the corresponding answer boxes (one letter for each possible answer) on a standard answer sheet bound in a test booklet.

When an airman finished the test, he noted the time on the front cover of the test booklet, handed it in, and left the room. Upon scanning the test booklets, it was discovered that all except one airman had cooperated in taking the test, leaving 30 airmen in the morning group and 31 airmen in the afternoon group. The test was then scored by Mr. J. E. Wilbourn and Mr. C. L. Cannon of the Personnel Research Laboratory and the resulting data was forwarded to The Shuford-Massengill Corporation for further analysis.

FUNDAMENTAL ANALYSIS OF THE DATA

The airmen went through the motions of taking a confidence test. They used the *SCoRule* to develop their answers and then they copied down the letters

from the answer boxes, thus, indicating various degrees of confidence in the different answers. Is there any meaning to the data that was produced?

What does it mean when an airman writes down the letter *A* for an answer implying that he has zero confidence that the answer is, in fact, the correct answer to the test item? What does it mean when he writes down a *Z* saying that he has complete confidence that the answer is the correct answer to the test item? What does it mean when he writes down an *M* implying that he has a confidence of .48, and so on for all the letters of the alphabet?

RESPONSE VALIDITY

One way of evaluating the meaning of this data is to examine for each letter of the alphabet the frequency with which it was used on a correct answer relative to the total frequency with which the letter was used. One would hope that the more confidence that an airman places on an answer, the more likely it is to be the correct answer. This does not necessarily have to be the case, however. An airman could use a random-like process for setting up the *SCoRule* and still produce a reasonable answer sheet with all sorts of different degrees of confidence indicated. If this process were totally at random within the constraints of the testing situation, then the expected relative frequency of an answer being correct would be about 20% and this would be so regardless of the particular confidence placed on the answer.

About 20% of those answers which have been assigned zero confidence (*A*) would be correct; 20% of those answers which have been assigned complete confidence (*Z*) would be correct; and so on for all possible degrees of confidence. In such a case, the data would have no meaning.

Tables 1a, b, c, d show these relative frequencies as a function of degree of confidence (indicated by the alphabet from *A* through *Z*) for each of the 61 airmen. For example, look at airman number *1* in Table 1a. Four hundred and thirty times he assigned *A* to an answer and 12 of those times he placed the *A* on a correct answer. The relative frequency of an answer to which this airman assigned zero confidence being correct is 12 divided by 430 or about .04. This airman placed an *M* on 21 of the answers and 9 of these were, in fact, correct answers giving a relative frequency of about .43. And finally, out of the 98 times that he placed *Z* on an answer, 87 of these answers were, in fact, correct yielding a relative frequency of about .89. Examination of the data in Table 1a, b, c, and d indicates that no airman followed such a random response strategy.

Table 2 shows this same data summed over all 61 airmen. The frequency with which the different degrees of confidence are used is a function of the difficulty level of the test. The relative frequency with which answers assigned given degrees of confidence are correct is not a function of the difficulty level of the test. Notice that, in general, this relative frequency increases for the higher degrees of confidence. This is shown more clearly in Figure 1 of the Appendix. There is clearly a functional relation between the degree of confidence assigned to an answer and the relative frequency with which it is, in fact, the correct answer. For the group as a whole, the more confidence an

airman places on an answer, the more likely it is to be the correct answer. In this sense, the responses of the group as a whole will certainly have meaning.

Does a similar relation hold for the data of each airman? The data of Tables 1a, b, c, and d suggest that maybe it does, but the variation due to the small number of observations for many of the degrees of confidence make it hard to see. A clearer picture may be obtained if one just looks at the relative frequency of an answer being correct when it was assigned an *A* and when it was assigned a *Z* since these are by far the most frequently occuring confidence assignments. These are shown for each of the 61 airmen in Table 3. In every instance, the Percent "Z" Answers Correct is larger than the Percent "A" Answers Correct. This suggests that each airman understood and did assign more confidence to the answers that he thought were correct. Thus, the data admits of a simple interpretation that the more confidence an airman assigns to an answer, the more likely it is to be correct.

GAIN IN INFORMATION

This alone, however, does not prove that one is getting completely meaningful data from confidence testing. To see this, one can consider a "blind" or "stupid" process which can yield a monotonic increasing relation (in fact it is a linear relation) between relative frequency and degree of confidence. Suppose that you were given the answer sheet of an airman who had taken AQE-66 as a choice test and all that has been indicated on the answer sheet is the number of the test item and which of the five answers the airman chose. You take this answer sheet and use it to fill out another answer sheet of the type used for confidence testing. Remember that all the items have five possible answers and that you must put down a letter indicating a degree of confidence for each of these answers. Say that on the first test item you look on the choice answer sheet to see which answer the airman has chosen and put a *Z* in this answer box indicating that the airman had complete confidence in that answer. Now, since all the confidence has been placed on that one answer, there is no confidence left over for the other answers, and an *A* must be placed in each of the remaining four answer boxes. You may continue doing this for a few more items but then change to a strategy of taking the answer chosen by the airman and giving it a confidence of about 1/2 (say either an *M* or an *N*) and giving one of the other answers a confidence of about 1/2. This leaves no confidence for the remaining three answers so give them all *A*'s. Follow this strategy for a while and then change to giving a confidence of about 1/3 to the answer chosen by the airman and a confidence of about 1/3 to each of two of the other answers leaving no confidence left over for the other two of the five answers. Now after a while, change to a strategy where you split the confidence four ways and then finally change to a strategy of assigning equal confidence to all of the five answers. Now you can vary back and forth between these strategies and mix them up any way you wish to get a reasonable appearing pseudo-confidence test answer sheet. You have not used any information that was not contained on the original choice answer sheet. You know nothing whatsoever about the test item, or whether or not the airman's answer was the correct one.

Now, what would this pseudo-confidence test data look like if we analyzed it as before. Suppose that the airman had chosen the correct answer 60% of the time in the original choice test. Now the pseudo-confidence test would show that about 60% of the *Z* answers were correct. For those items where the confidence was split between two answers, the percent of *M* answers correct would be 1/2 of 60 plus 10 equals 70 or 35%. For those items where the confidence was split between three answers, the percent of *I* answers correct would be 1/3 of 60 plus 10 plus 10 equals 80 or 26-2/3%. For those items in which the confidence was divided among four answers, the percent of *G* answers correct would be 60 plus 10 plus 10 plus 10 equals 90 or 22-1/2%, while for those items in which the confidence was divided among all five answers, the percent of *F* answers correct would, of course, be about 20%. Then finally, the percent of *A* answers correct would be about 10%. If this relation were plotted on a graph, the expected values would fall on a straight line and, of course, the observed values would vary randomly around this straight line.

The main thing, however, is that there would be a functional relation between relative frequency and degree of confidence, but this relation is certainly not one that we would be happy with. In generating the data that yielded this relation, you used no information other than contained in the choice responses of the airman. Therefore, the pseudo-confidence responses can reflect no valid information other than that contained in the choice responses. In order for confidence responses to contain valid information over and above that yielded by choice tests, the relation between relative frequency and degree of confidence *must be steeper than* this base-level relation which depends upon the percentage of correct answers in the choice administration of the test.

Therefore, in order to determine whether or not any additional information has been obtained from the confidence measurement administration of AQE-66, the percent of *Z* answer correct must be compared with the percent of correct answers that would have been obtained if the test had been administered as a choice test. A confidence test can always be scored as a choice test if you are willing to make the assumption that the examinee would have chosen that answer in which he indicated the greatest degree of confidence. In the event that two or more answers are approximately tied for maximum degree of confidence, then the tie can be broken by using a table of random numbers. This was exactly the procedure used to obtain the inferred percent of correct answers shown in Table 3. Notice that in every case the percent of *Z* answers correct exceeds the inferred percent correct answers. These comparisons are also shown graphically in Figure 1' of the Appendix. It is evident that there are great individual differences in these data. These differences in part reflect an airman's ability to evaluate information in terms of how much confidence is justified by the information at hand. For example, airman number 26 as shown by the data of Table 1b and Table 3 does not know what he knows and what he doesn't know. He sees things in terms of black and white and his data is very much like that yielded by a choice test. In fact, it appears that the worst that confidence measurement can do is to yield data like that from a choice test. Airman 25, on the other hand, is exceptionally good at evaluating information. He evidenced many degrees of confidence on different test items and his percent *Z* answers correct is about 92 and his inferred percent correct answers is about 55.

DISRUPTION OF THE TESTING PROCESS

So far the analysis of response validity has been internal to the experimental readministration of AQE-66 as a confidence test. By taking account of information external to this test administration we can evaluate another hypothesis about confidence testing. As mentioned before, there has been some concern that the procedures of confidence testing are so complex as to confuse some of the airmen and thus, lower their test performance. Confidence testing would so interfere with the test-taking process that a distorted picture would be obtained of the airman's ability level. This is certainly a conceivable outcome and if it happened, we should find that the Percent "Z" Answers Correct computed for some airmen were actually lower than the Percent Correct Answers yielded by the same airmen during an independent administration of the same test as a choice test. We did not find this happening with respect to the Inferred Percent Correct Answers which is one indicant of a score which a student would have obtained if he had simultaneously had been given a choice test. This is no real test of the hypothesis, however, since the Inferred Percent Correct Answers is derived from the confidence test data and if it were distorted by lack of understanding of the student, then the distortion should occur both in the Percent "Z" Answers Correct and in the Inferred Percent Correct Answers.

As mentioned above, these airmen had already taken AQE-66 prior to entering service with the Air Force and, thus, their original test performance could provide an independent but somewhat remote check on extent to which the airmen were confused by confidence testing. We were able to obtain the original test data of only 40 out of the 61 airmen. The percent of correct answers given by these 40 airmen during the original administration of AQE-66 is shown also in Table 3. In every case but three, the airmen's Percent "Z" Answers Correct is greater than the Original Percent Correct Answers. The three exceptions are airmen number *7*, *10*, and *56* with the greatest deviation being represented by airman number *10*. When this airman took AQE-66 for the first time and as a choice test he got 70% of the answers correct. When he took it over again as a confidence test, about 60% of the answers in which he had complete confidence were, in fact, correct answers. This is a difference of about .10. It is possible that this airman was confused by the procedures of confidence testing but it should be noted, however, that his Inferred Percent of Correct Answers was about 53% which is somewhat below the 60% "Z" Answers Correct indicating that his test data yielded information over and above that that would have been obtained from a choice administration. On the other hand, airman number *56* got about 73% correct answers in the original administration of AQE-66, but we infer that he would have gotten about 45% correct in the readministration of the test. This is a considerable drop. It should be noticed that he did get about 70% correct when he said he had complete confidence in an answer. This indicates a considerable gain for airman number *56* in the information obtained from confidence testing over that obtained from choice testing. All in all, there is very little evidence that the procedures of confidence testing interferred with the test-taking performance of these airmen. There is no doubt that it can happen, but with careful instruction in the procedures of confidence testing the relative frequency of occurrence of confusion in examinees should be reduced to near zero.

It is unfair to an examinee if he fails to clearly understand instructions for

taking a test and if he fails to adopt a test-taking strategy which maximizes his expected test score for the amount of knowledge that he possesses and this will be the case if an airman were confused by the procedures of confidence testing. It is also the case, however, if an airman fails to respond to all the items in the choice administration of AQE-66 since his expected test score is maximized if and only if he responds to all items even to the extent of guessing in those situations where he does not know the answer. This happens and has been shown to have major effect on the fairness of a choice test (Shuford and Massengill, 1963).

THE EXISTENCE OF GUESSING

The possibility of guessing in multiple-choice tests has been recognized as a problem. In any highly developed and perfected test such as AQE-66, several techniques have been used to minimize the existence of guessing. First, the decision to use five alternative multiple-choice items is intended to minimize the effect of guessing on the test results. Second, a major goal in the writing of test items and the possible answers is to write them in such a way that when a person doesn't know the correct answer, he will be almost certain to pick out one of the misleads. In the language of confidence testing the goal is to write test items so that an examinee either has a very high degree of confidence in a correct answer or a very high degree of confidence in one of the incorrect answers. If this goal were achieved, it would certainly minimize the effect of guessing, since guessing occurs only when the examinee is uncertain between the correct answer and one or more of the incorrect answers. For example, an airman would be in a guessing situation if he assigned a confidence of about 1/2 to the correct answer and a confidence of 1/2 to one of the incorrect answers. He would have enough partial information to rule out three of the incorrect answers but he would still be undecided between a correct answer and one of the incorrect answers. If he flipped a "mental coin" to decide between these two answers his probability of getting the right answer would be 1/2. On other items, the airman could be undecided between three out of five having partial information to rule out two of the incorrect answers and a probability of chance success of 1/3; undecided between four out of the five having enough partial information to rule out one of the incorrect answers and a probability of chance success of 1/4; and finally, he could be totally uninformed and have equal confidence on all of the five answers with a probability of chance success of 1/5.

Now the Airman's Qualifying Examination has undergone a great deal of development and refinement over years. Empirical large scale item analysis procedures have been used to select the items used in AQE-66 from large pools of available items. In this sense then, AQE-66 represents a highly refined test where a great deal of effort has been devoted to eliminating the effect of guessing on the test results. In spite of this, the confidence responses of the airmen indicated that they were encountering guessing situations on about 1/4 of the items. This is a major and highly promising finding because every time an examinee encounters a guessing situation in taking a test, his response is contributing error variance to the test results and this error variance, of course, reduces the reliability, validity, and efficiency of the test data (Shuford and Massengill, 1967b). In confidence testing, however, these guess-

ing situations are discriminated from other states of knowledge and thus, do not add to the error variance of the test results. Thus, to the extent that guessing is a factor in choice testing there is corresponding room for improvement by administering the test as a confidence test rather than a choice test. Since guessing situations are encountered by the airmen taking AQE-66, let's see what effect this has on derived measures such as total scores and aptitude scores for AQE-66.

PSYCHOMETRIC ANALYSIS OF CHOICE AND CONFIDENCE DATA

AQE-66 is made up of ten subtests (Vitola and Madden, 1967). Nine of the subtests are composed of multiple-choice items and together make up the 150 multiple-choice items analyzed in the previous section. The tenth subtest is composed of 60 computational items and is not included in this analysis. The scores from the subtests are added together in various ways to derive four different aptitude scores called *General*, *Administrative*, *Mechanical*, and *Electronics*. Table 4 shows the number of items that make up each one of these aptitude scores. There is, of course, some overlap between the aptitude scores in the sense that item scores enter into both aptitude scores.

Table 4 also shows the corresponding means and standard deviations of the aptitude and total test scores for the original administration of AQE-66 and for the experimental readministration both for the *Valid Confidence* score and for the total amount of confidence assigned to the correct answers. Notice that the confidence scores are higher than the other mean scores. This reflects the nonlinear characteristic of the admissible scoring system used in *Valid Confidence* Testing. Notice also that the standard deviations for the *Valid Confidence* scores are smaller than in the other cases. This again is a reflection of the logarithmic admissible scoring system. The scores of the poorest performing airmen are being raised considerably and the whole distribution of scores is being compressed toward the upper end of the scale. In this sense then, AQE-66 is probably too "easy" for optimal performance as a confidence test.

RELIABILITY

Both theory (Shuford and Massengill, 1966b; 1967b) and intuition suggests that eliminating error variance due to guessing will increase the reliability of a test. Table 5 shows the published reliabilities (Vitola and Madden, 1967) for the four aptitude indexes of AQE-66. These published reliabilities include the score from Arithmetic Computation Subtest which enters only into the Administrative Aptitude score. Thus, this reliability is not exactly comparable to the reliabilities we compute. Table 5 also shows the reliability indexes computed on the choice data of the 40 airmen for whom we could obtain the original answer sheets. Allowing for variation due to sampling, these reliabilities seem to be in line with the published figures. Notice that the reliability index for all 150 items is .954. AQE-66 is, quite obviously, an exceptionally reliable test.

Table 5 shows also the reliability indexes computed for the *Valid Confidence*

scores of the 40 airmen. Reliabilities of the *Valid Confidence* scores are higher than the reliabilities of the original choice scores, but not much higher. As mentioned above, the variances of the *Valid Confidence* scores are considerably smaller than those of the choice scores. Ordinarily such a reduction in variance results in a reduction in the size of a correlation coefficient such as the reliability index. This did not happen in this case. The variances got smaller but the correlation coefficient increased in size.

We can get a better picture of the gain in reliability obtained from confidence testing if we look at the total amount of confidence assigned to correct answers. This measure yields score distributions much more comparable both in terms of means and variances to the distributions of the choice scores from the original administration. The reliability indexes for this measure are also shown in Table 5 and in every case exceed all other reliability indexes in size.

Although these differences in reliability may appear to be trivial, they are not. The correlation coefficient is not a linear measure of testing efficiency. As the correlation coefficient approaches 1, smaller and smaller differences become more important.

One way to evaluate the importance of the gain in reliability resulting from changing from choice administration to confidence administration of AQE-66 is in terms of test length. Consider for example how many additional items would be required to make the choice test as reliable as the confidence test according to the total amount of confidence assigned to the correct answers. Table 6 gives the answer as derived from the Spearman-Brown Prophecy Formula. From 37 up to 56 additional items will have to be added to parts of AQE-66 to make each aptitude score as reliable as that obtained from giving the current AQE-66 as a confidence test. For the test as a whole, 121 items will need to be added to AQE-66. In a sense then, administering AQE-66 as a confidence test has the effect, in terms of reliability, equivalent to considerably increasing the number of items in the test.

Another way of looking at the relative efficiency of choice and confidence testing is to consider how many items can be eliminated from the confidence version of AQE-66 to reduce its reliability down to that of the choice version of AQE-66. The answer, also derived from the Spearman-Brown Prophecy Formula is shown in Table 6. From 20 to 29 items can be eliminated from those parts making up each aptitude score and for the test as a whole, 66 items could be eliminated. In other words, an AQE-66 consisting of only 84 items administered as a confidence test would be just as reliable as the present AQE-66 of 150 items administered as a choice test. Savings indicated in Table 6 are probably underestimates of what actually can be achieved in practice since the projections are based upon the assumption of blind random choice of items to be omitted. If item analysis information were used to make an optimal selection of items for inclusion in the test, then it should be possible to make the test even shorter than indicated.

INTERCORRELATION BETWEEN APTITUDE SCORES

Eliminating error variance due to guessing from AQE-66 should not only improve the reliability of the test, but should, at least under some conditions, in-

crease the correlation between the aptitude scores. While increasing reliability is a desirable effect, increasing intrabattery correlations reduces the ability of the test to make differential predictions. Given whatever "true" correlation there might be between the aptitude scores, the introduction of random error due to guessing would serve to lower the computed correlations but could not improve differential prediction. If this were happening, then eliminating the random error due to guessing would allow the higher "true" correlations to manifest themselves.

Table 7 shows the correlation between pairs of aptitude scores for AQE-66. The first column shows the published figures while the second column shows the correlations computed from the original choice test data of our 40 airmen. The greatest deviation between the published figures and those obtained from the small sample of 40 is for the correlations between Mechanical and Electronic Aptitude scores where the correlation based upon the small sample is quite a bit less than that in the large-sample published data. It should be noted that three of the six correlations are not exactly comparable to the published data because the published data includes the arithmetic computation subtest as a component of its Administrative Aptitude score.

Table 7 also shows the correlations based on *Valid Confidence* scores and the total amount of confidence assigned to correct answers computed from the data of the 40 airmen. In every instance these correlations are at least as large as the correlations computed from the original choice data of the same 40 airmen. Figures 2, 3, and 4 of the Appendix graphically display the relations between these pairs of aptitude scores.

PREDICTIVE VALIDITY OF CHOICE AND CONFIDENCE DATA

The increased reliability and intrabattery correlations obtained in the confidence administration of AQE-66 could be due, in whole or in part, to the reduction in error variance produced by the guessing occuring when AQE-66 is administered as a choice test. It could also be produced, in part, by confidence testing introducing new variance which is not present when AQE-66 is administrated as a choice test. This must happen to a certain extent. Remember that there are wide individual differences in the airman's ability to evaluate information. This is one reason why confidence testing yields a different rank order of examinees and would most certainly be operating pretty much throughout the test and effecting each aptitude score. Although there are reasons for accounting for this as true variance and believing that it should be measured and reflected in any test score, an unequivocal answer cannot be obtained without further research and validation studies.

There is, however, one bit of data that can be used to get some hint at the ability of confidence testing to improve the validity of AQE-66. The records of the airmen used in this study contain their Air Force Qualifying Test scores (AFQT). This is, like AQE-66, a multiple-choice aptitude test. If these Air Force Qualifying Test scores were considered as a criterion variable, the correlation between AQE-66 and AFQT would indicate the validity of AQE-66 considered as a predictor of AFQT. Figure 5 in the Appendix shows the relation between AQE-66 administered as a choice test and AFQT for the 40 airmen. The correlation between these two sets of scores is .76.

Figure 6 in the Appendix shows the relation between the *Valid Confidence* score for AQE-66 administered as a confidence test and AFQT. The reduction in variance produced by the logarithmic admissible scoring system is quite apparent here. Correlation for these two sets of scores is .70.

When we attempt to equalize variances, however, by looking at the total amount of confidence on correct answers obtained from AQE-66 administered as a confidence test, we find the relation with AFQT shown in Figure 7 of the Appendix. The AQE-66 data is now much more spread out with a variance comparable to that obtained from the choice administration of AQE-66. The correlation in this case is .81 which is somewhat above that of .76 found for the relation between AQE-66 administered as a choice test and AFQT. Thus, the administration of AQE-66 as a confidence test can increase its predictive validity.

DISCUSSION AND CONCLUSIONS

The results of this study do not prove that confidence testing yielded improvements in personnel selection, classification and placement. It would be unreasonable to expect this from such an experiment. There is, on the other hand, nothing in these data to deny the possibility that confidence testing can radically improve the testing process.

This small scale study may be viewed as setting up a series of hurdles for confidence testing to pass, and pass them it did. First, the airmen did not get marred down in a sea of confusion about the procedures of confidence testing. In general, they responded with remarkable intelligence and yielded test data which contains information over and above that which is possible to get when AQE-66 is administered as a choice test.

These data suggest that there are wide individual differences in the way in which airmen process and evaluate information. These differences are reflected in their test scores, but until there is reason to believe that this is not characteristic of their behavior outside of this particular test situation there should be no cause for concern. In fact, this may prove to be a new source of true variance which would serve to improve the validity of test data.

There was no dearth of guessing situations encountered by these airmen in responding to AQE-66 even though it is a highly refined test designed to minimize guessing. This gave confidence testing an opportunity to reduce one source of error variance in AQE-66 and this was reflected in the higher reliability and intrabattery correlations and in the improved correlation between AQE-66 and AFQT.

REFERENCES

de Finetti, B. Does it make sense to speak of good probability appraisers? In I. J. Good (Gen. Ed.) *The scientist speculates.* New York: Basic Books, 1962. Pp. 357-364.

Massengill, H. E. & Shuford, E. H. (1966) *Decision-theoretic psychometrics: a logical analysis of guessing.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Massengill, H. E. & Shuford, E. H. (1967) *What pupils and teachers should know about guessing.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Massengill, H. E. & Shuford, E. H. (1968) *A report on the effect of degree of confidence in student testing.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Roby, T. B. (1965) *Belief states: a preliminary empirical study.* ESD-TDR-64-238, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Massachusetts.

Shuford, E. H., Albert, A. & Massengill, H. E. (1966) Admissible probability measurement procedures. *Psychometrika,* 31, 125-145.

Shuford, E. H. & Massengill, H. E. (1966a) *Decision-theoretic psychometrics: the effect of guessing on the quality of personnel and counseling decisions.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, E. H. & Massengill, H. E. (1966b) *Decision-theoretic psychometrics: the worth of individualizing instruction.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, E. H. & Massengill, H. E. (1967a) *The relative effectiveness of five instructional strategies.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, E. H. & Massengill, H. E. (1967b) *How to shorten a test and increase its reliability and validity.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Shuford, E. H. & Massengill, H. E. (1967c) *Individual and social justice in objective testing.* Lexington, Massachusetts: The Shuford-Massengill Corporation.

Toda, M. (1963) *Measurement of subjective probability distributions.* ESD-TDR-63-407, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Massachusetts.

van Naerssen, R. F. (1961) A scale for the measurement of subjective probability. *Acta Psychologica,* 159-166.

Vitola, B. M. & Madden, H. L. (1967) *Development and standardization of airman qualifying examination-66.* Personnel Research Laboratory, Lackland Air Force Base, Texas

Table 1a. Response Frequency as a Function of Degree of Confidence. Right hand number of entry shows frequency with which degree of confidence was used while left hand entry shows frequency with which that degree of confidence was assigned to a correct answer.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a	i	12/430	37/461	14/341	12/492	5/497	16/297	54/283	19/405	16/520	50/446	39/342	11/431	37/236	50/386	12/594
b	i							2/6						1/2	1/7	
c	i					0/1	1/1	2/10			0/2	2/3	0/1		0/1	
d	i	1/8			1/4	1/4	1/7	7/32	0/2		3/15	3/15			1/3	
e	i			24/101	1/5	0/9	19/78	16/70			3/21	14/50	2/12	17/87	8/31	
f	i	32/162	21/105	18/86	3/15	16/83	21/120	24/128	25/123	14/70	9/33	14/108	34/169	27/147	13/59	4/20
g	i	4/16	5/22	18/104	4/16	2/5	14/66	14/71	11/48	2/5	5/23	11/43	0/1	24/99	15/76	
h	i	0/1	0/1	1/9		0/1	4/26	7/34			5/18	6/17		1/9	7/41	
i	i	1/3	5/13	2/7	4/12	1/4	8/22	4/20	10/43	2/6	0/8	2/10	1/3	4/16	3/17	
j	i			0/3		0/2	2/10	0/6	0/2		1/9	3/10	1/3	1/8	0/2	
k	i			0/2		1/1	1/2	3/8	1/2	1/1	1/8	2/7	1/5	0/3	1/7	
l	i	1/1		1/2		0/1		0/3	0/1		0/2	3/7		2/3	2/4	
m	i	9/21	8/44	3/4	7/17	2/6	12/27	1/8	17/38	6/19	3/11	7/28	7/14	5/20	2/13	0/2
n	i			0/2	0/2	1/1	0/3	0/1			1/2	4/15		1/16	1/6	0/4
o	i			2/3		0/2		0/2			0/2	1/4	1/1	1/1	0/2	
p	i					1/2			1/1		0/1		1/1		0/4	
q	i					3/3			1/2		1/2	2/4			0/1	
r	i					2/2					1/1	0/2	1/1			
s	i	1/1				2/3				0/1				1/1		
t	i					4/4							1/1			
u	i					1/2										
v	i	2/2														
w	i					0/1										
x	i															
y	i						1/1								0/1	
z	i	87/98	74/97	67/79	105/114	108/111	44/53	16/61	65/76	109/121	59/99	33/58	89/100	17/40	46/84	131/143

Table 1b. Response Frequency as a Function of Degree of Confidence. Right hand number of entry shows frequency with which degree of confidence was used while left hand entry shows frequency with which that degree of confidence was assigned to a correct answer.

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
a i	25/518	10/399	10/505	20/462	17/538	17/537	43/454	16/298	12/372	10/336	105/563	14/382	12/541	7/293	15/417
b i		1/10		0/11						0/1		0/1		0/4	0/5
c i		3/8		2/10					0/8	0/4		0/6	0/3	3/17	0/4
d i	0/2	2/5		0/3		0/2	3/5		0/4	2/14	0/1	1/6	0/4	9/71	1/2
e i	4/14	1/9	1/3	4/13	0/1	1/3	0/4		47/235	2/8		0/4	1/12	12/55	4/33
f i	5/24	27/136	4/22	14/71	7/32	5/35	24/117	11/55	0/6	12/60		14/62	4/22	19/91	20/95
g i	3/9	4/14	1/5	4/19	1/4	6/24	1/3	52/243		0/9		4/20		15/78	10/42
h i	2/4	3/5	1/2	5/12	2/6		2/5	5/23	0/1	3/10		1/5	0/1	3/13	1/6
i i	0/2	1/7	1/5	1/4	2/13		1/7	9/29	1/4	4/13		3/6	2/5	0/7	2/12
j i	0/1	1/3	0/5	0/1	2/6		1/6		0/1	2/9		0/6	3/6	0/4	3/6
k i	0/1	0/1	1/1	2/4	0/2	1/3	1/5		0/1	1/1			2/2	2/2	
l i	2/4	0/1		0/3	2/3		1/8		3/4			0/4	1/3	0/2	
m i	15/37	0/1	2/11	3/8	3/5		2/9	6/18	1/4	5/10	10/61	4/14	4/7	8/15	5/16
n i	3/8	2/4	1/1	1/2	0/4	1/2	7/13	3/7	2/4	1/5	2/2	2/5	2/6	6/13	1/2
o i	0/2	0/2	1/1	3/5	1/2		1/6		1/2	1/2		2/3	0/1	3/6	1/4
p i	1/1	1/3	0/1			2/3	2/3		1/2	1/1		1/1	0/1	0/2	1/2
q i	3/4		3/3						1/2	5/5		3/3			
r i	0/2	1/1	1/1	2/4					4/4	3/4		1/2	1/1		1/1
s i	0/1		1/1	0/3					1/1	1/3				2/2	
t i	0/3					0/1			2/2	1/2		0/1			1/1
u i	1/2	1/1	2/2	3/3		1/1				2/5		0/1		1/1	1/1
v i	2/2	0/1		1/1					3/4	1/1		2/3		1/1	
w i		0/1		0/2		1/2			1/1			2/2		1/1	2/3
x i												2/3			2/2
y i		1/1						1/1							
z i	84/102	84/94	108/114	85/101	113/127	115/130	61/97	47/59	70/81	57/62	33/117	68/74	118/128	58/64	79/89

Table 1c. Response Frequency as a Function of Degree of Confidence. Right hand number of entry shows frequency with which degree of confidence was used while left hand entry shows frequency with which that degree of confidence was assigned to a correct answer.

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
a i	8/590	38/366	18/532	9/400	49/521	45/483	37/485	22/483	35/451	41/358	12/242	19/538	75/390	10/505	29/346
b i						0/2		0/1	0/3	1/4	1/12		2/11		
c i			1/4			2/5		0/1	1/3	1/5	4/17		2/9		
d i					1/1	0/15	0/1	0/1	1/10	4/8	6/34	1/3	0/4	0/1	1/1
e i				38/188		7/16		1/10	8/29	2/13	12/73	0/3	2/16	6/22	0/1
f i		16/78	14/70	3/15		4/24	20/97	1/10	16/84	38/185	32/149	2/11	1/15	12/60	8/38
g i	1/4	1/4		3/13	1/2	2/20	1/2	2/12	1/12	5/24	17/67	1/3	8/34	2/20	2/7
h i		0/2		2/8	1/4	1/14		0/2	1/5	1/14	5/27	0/1	6/20		
i i		9/32			0/5	1/6		3/6	2/5	9/29	3/13		2/8		4/17
j i		2/8	1/1		1/5	1/1	3/13	1/2	2/5	0/1	5/8		4/12	0/1	0/1
k i				1/3	2/3	2/10			0/2	0/2	2/9	0/1	3/14	1/1	
l i					0/6	5/12		1/1	3/4	1/1	1/2	0/1	3/18	0/1	1/1
m i		28/74		10/27	12/40	2/11	20/47	12/29	7/14	6/14	2/9	2/6	6/31	2/3	2/6
n i		0/2			4/18	3/9		0/3	1/12		1/1	2/3	2/20	1/3	2/3
o i					0/2	2/8		0/2	0/1	0/3	1/4		6/20		0/1
p i					0/2	2/5		1/1			0/1	1/1	0/2		0/1
q i			0/1		1/1				0/1	0/1	1/3				1/2
r i						1/4		1/1	0/1		0/4		0/1		
s i						1/3					0/1	0/2	0/1		0/1
t i			0/1		0/1	1/1		2/2	0/1		1/4	0/1		1/1	1/1
u i						0/1		2/2		0/1	0/1	3/4		1/1	1/2
v i						1/2				0/1				1/1	1/1
w i								1/1		0/1		2/2			
x i													0/2		
y i		9/11											1/1	1/1	
z i	141/149	28/51	116/134	84/89	72/102	67/91	69/98	86/103	72/100	41/78	41/47	110/128	16/59	112/122	49/75

Table 1d. Response Frequency as a Function of Degree of Confidence. Right hand number of entry shows frequency with which degree of confidence was used while left hand entry shows frequency with which that degree of confidence was assigned to a correct answer.

		46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
a	i	10/308	19/525	30/411	40/526	3/597	9/375	22/362	30/430	19/494	49/400	24/323	40/496	19/238	15/523	16/234	10/516
b	i	0/3			0/1			3/18	2/2	1/4		0/17		1/2			0/3
c	i	0/6	0/1					1/13	0/1	1/3		2/7		0/4			0/1
d	i	2/14	2/6	1/9			3/18	2/22	6/24	1/7		2/13		7/27	1/2		0/1
e	i	3/15	4/10	7/35	3/8		18/94	7/27	2/13	0/9	0/3	4/15		4/41	3/4		
f	i	51/251	1/10	4/15	2/12		19/85	4/20	4/27	10/51	35/151	2/11	19/94	16/69	2/5	62/307	6/31
g	i	9/33	3/25	8/34	1/6		7/33	11/52	6/21	5/10	12/68	2/10	0/2	12/74	2/9		0/1
h	i	2/7	3/8	8/18	2/6		1/13	9/29	4/33	3/5	0/10	0/2	3/6	17/54	1/4		
i	i	2/10	1/4	2/25			5/12	5/20	9/23	1/6	0/2	0/2	1/8	5/25	3/10	1/3	
j	i	0/4	3/4	2/4			1/3	1/11	5/24	1/3		1/4	2/3	6/20	0/3		2/9
k	i		0/4	3/11				4/18	4/20	0/1	0/2	2/4	0/2	2/5	0/1		0/1
l	i	0/1	1/4	5/14	1/3			6/14	2/12	0/1	1/1	1/3	1/3	0/8	1/5		
m	i	4/11	7/14	9/35	4/21	1/1	7/18	8/23	3/6	11/23	3/8	4/12	3/7	2/11	2/5	7/26	14/30
n	i	6/9	1/3	11/31	3/10	0/1	5/11	6/17	1/9	1/2		1/3	2/5	2/4	2/5		1/1
o	i	2/2		2/7				0/14	1/5	0/1	0/1	1/1	0/2	1/4	1/2		
p	i			3/4				2/6	4/8	3/4		3/3	0/1	1/1			
q	i							3/4	1/1	2/3		0/1		1/1	1/2		
r	i		1/3	0/1	1/2			1/1	2/4	0/2		0/2		1/1	1/2		1/1
s	i		2/4		0/1		1/2	1/1	2/2	1/2		3/6		2/3	1/3		
t	i	1/2	3/5	1/1	1/1					3/5		2/3			1/2		
u	i	1/1		0/1			1/1	1/1	1/2	2/4				2/2			
v	i	2/2	1/2					1/1	1/2	1/1		1/1			0/1		2/2
w	i	2/2	1/1					1/2	1/3	0/1							
x	i	1/1	1/1					1/2		0/1		1/1					
y	i											2/2			0/1		
z	i	52/61	96/109	48/57	86/115	146/149	73/78	50/56	59/71	84/99	50/97	45/64	79/115	26/35	105/114	39/49	108/116

Table 2. Response Frequency and Relative Frequency of Answer being correct as a Function of Degree of Confidence.
Entries based on all 61 airmen. Right hand number of entry shows frequency with which degree of confidence was used while left hand entry shows frequency with which that degree of confidence was assigned to a correct answer.

Degree of Confidence	Response Frequency	Relative Frequency
A	1,519/26,243	.058
B	16/130	.123
C	28/159	.176
D	77/431	.179
E	312/1,503	.208
F	845/4,240	.199
G	355/1,664	.213
H	129/542	.238
I	142/563	.252
J	64/261	.245
K	48/183	.268
L	52/172	.302
M	357/1,056	.338
N	100/315	.317
O	36/133	.271
P	33/69	.478
Q	33/50	.660
R	28/56	.580
S	23/49	.469
T	27/47	.574
U	28/43	.651
V	24/32	.750
W	15/26	.577
X	8/13	.615
Y	16/18	.889
Z	4,481/5,584	.802

Table 3. Individual Response Validity Data Compared with Choice Data.

AIRMAN	PERCENT "A" ANSWERS CORRECT*	PERCENT "Z" ANSWERS CORRECT*	INFERRED PERCENT CORRECT ANSWERS*	ORIGINAL PERCENT CORRECT ANSWERS
1	2.8	88.8	68.0	---
2	8.0	76.3	55.3	---
3	4.1	84.8	52.7	71.3
4	2.4	92.1	73.3	46.5
5	1.0	97.3	85.3	80.7
6	5.4	83.0	43.3	39.3
7	19.1	26.2	18.7	33.3
8	4.7	85.5	60.0	60.7
9	3.1	90.1	78.7	76.7
10	11.2	59.6	53.3	69.7
11	11.4	56.9	38.0	42.7
12	2.6	89.0	70.0	58.7
13	15.7	42.5	23.3	40.3
14	13.0	54.8	40.7	39.3
15	2.0	91.6	88.7	86.0
16	4.8	82.4	70.7	50.7
17	2.5	89.4	65.3	51.1
18	2.0	94.7	79.3	73.3
19	4.3	84.2	70.7	66.0
20	3.2	89.0	78.0	71.3
21	3.2	88.5	82.7	78.7
22	9.5	62.9	51.3	53.3
23	5.4	79.7	46.7	56.7
24	3.2	86.4	62.7	---
25	3.0	91.9	55.3	56.7
26	18.6	28.2	28.0	---
27	3.7	91.9	56.7	---
28	2.2	92.2	85.3	---
29	2.4	90.6	57.3	---
30	3.6	88.8	71.3	---

Table 4. Number of Items, Means, and Standard Deviations for Original and Experimental Administration of AQE--86.
Statistics based on the 40 airmen for whom original test data could be retrieved.

	GENERAL APTITUDE	ADMINISTRATIVE APTITUDE	MECHANICAL APTITUDE	ELECTRONICS APTITUDE	ALL TEST ITEMS
Number of Items	59	46	62	61	150
Original Administration					
Number of Correct Answers					
Mean	41.7	32.9	35.7	33.2	92.1
Standard Deviation	10.86	8.67	10.38	11.14	24.21
Experimental Administration					
Valid Confidence Score					
Mean	46.2	34.6	43.8	40.4	108.2
Standard Deviation	9.51	7.86	7.95	9.17	20.71
Total Amount of Confidence Assigned to Correct Answer					
Mean	40.0	31.0	36.0	33.8	91.4
Standard Deviation	12.34	9.85	11.08	11.85	27.05

AIRMAN	PERCENT "A" ANSWERS CORRECT*	PERCENT "Z" ANSWERS CORRECT*	INFERRED PERCENT CORRECT ANSWERS*	ORIGINAL PERCENT CORRECT ANSWERS*
31	1.4	94.6	94.0	94.0
32	9.8	54.9	43.3	41.3
33	3.4	86.6	80.0	78.0
34	2.2	94.4	66.0	69.3
35	9.4	70.6	53.3	68.7
36	9.3	73.6	57.3	52.6
37	7.6	70.4	58.0	---
38	4.6	83.5	68.7	76.0
39	7.8	72.0	52.7	---
40	11.4	52.6	36.0	---
41	5.0	87.2	45.3	---
42	3.5	85.9	80.7	81.3
43	19.2	27.1	24.7	---
44	2.0	91.8	74.0	---
45	8.4	65.3	40.7	36.9
46	3.2	85.2	54.7	---
47	3.6	88.1	75.3	---
48	7.3	84.2	47.3	59.3
49	7.6	74.8	62.0	53.3
50	.5	98.0	97.3	94.0
51	2.4	93.6	66.0	63.3
52	6.1	89.3	56.0	70.7
53	7.0	83.1	56.7	---
54	3.8	84.8	71.3	70.0
55	12.2	51.5	37.3	36.7
56	7.4	70.3	44.7	72.7
57	8.1	68.7	61.3	52.0
58	8.0	74.3	40.0	---
59	2.9	92.1	76.0	---
60	6.8	79.6	36.7	---
61	3.0	93.1	79.3	---

Table 5. Reliability Indices (KR–20) for Original and Experimental Administration of AQE–66.
Statistics based on the 40 airmen for whom original test data could be retrieved.

	GENERAL APTITUDE	ADMINISTRATIVE APTITUDE	MECHANICAL APTITUDE	ELECTRONICS APTITUDE	ALL TEST ITEMS
Number of Items	59	46	62	61	150
Published Reliabilities (KR-21)	.94	.93	.88	.84	--
Original Administration					
Number of Correct Answers	.915	.910	.913	.890	.954
Experimental Administration					
Valid Confidence Score	.919	.917	.931	.892	.961
Total Amount of Confidence Assigned to Correct Answers	.948	.948	.952	.935	.974

Table 6. Comparison of the Efficiencies of Choice and Confidence Testing for AQE—66 based on Total Amount of Confidence Assigned to Correct Answers.

	Number of Additional Items Required to Make Choice Test as Reliable as Confidence Test	Number of Items that Must Be Discarded to Reduce Reliability of Confidence Test to that of Choice Test.
General Aptitude	56	-29
Administrative Aptitude	37	-20
Mechanical Aptitude	48	-26
Electronics Aptitude	41	-24
All Test Items	121	-66

Table 7. Correlations between Pairs of Aptitude Scores for AQE-66. Statistics based on the 40 airmen for whom original test data could be retrieved.

	Published for AQE-66	Original Administration Choice Score	Experimental Administration	
			Valid Confidence Score	Total Amount of Confidence Assigned to Correct Answers
General-Administrative	.82	.93*	.96*	.95*
General-Mechanical	.79	.77	.83	.78
General-Electronics	.83	.79	.91	.87
Administrative-Mechanical	.59	.65*	.72*	.65*
Administrative-Electronics	.79	.80*	.93*	.89*
Mechanical-Electronics	.82	.67	.78	.76

*Administrative Score lacks Arithmetic Computation Subtest.

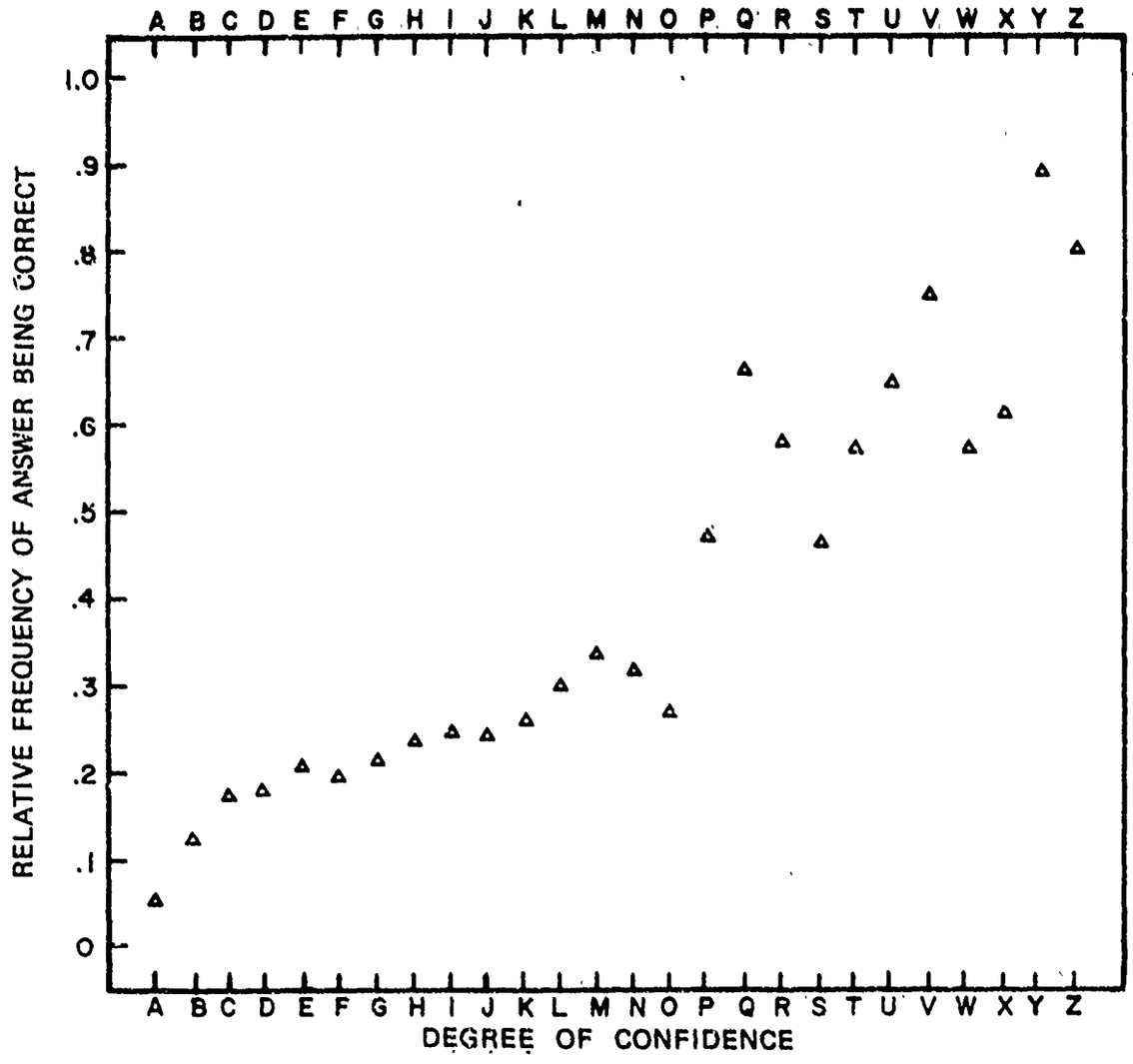


Figure 1. Response validity function based on all 61 airmen.

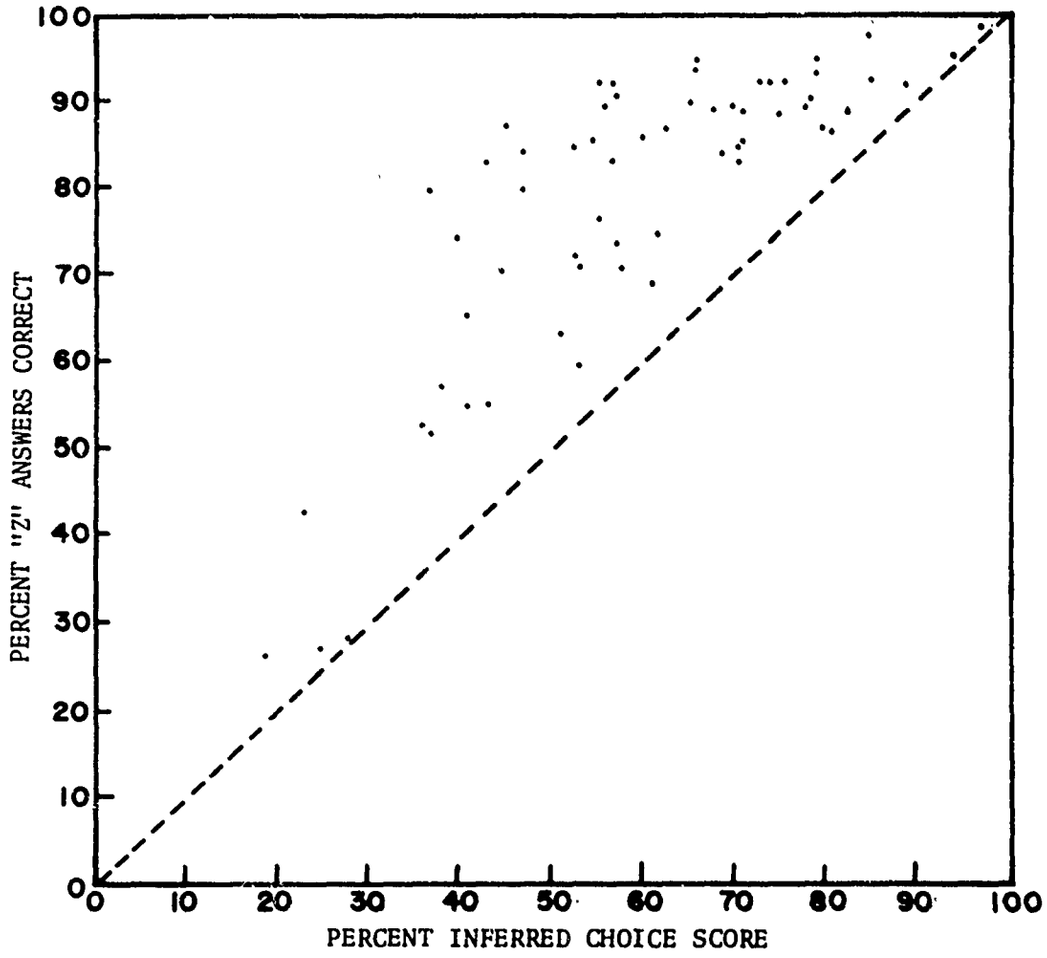


Figure 1'. Gain in information from confidence testing. Data shown for each of the 61 airmen.

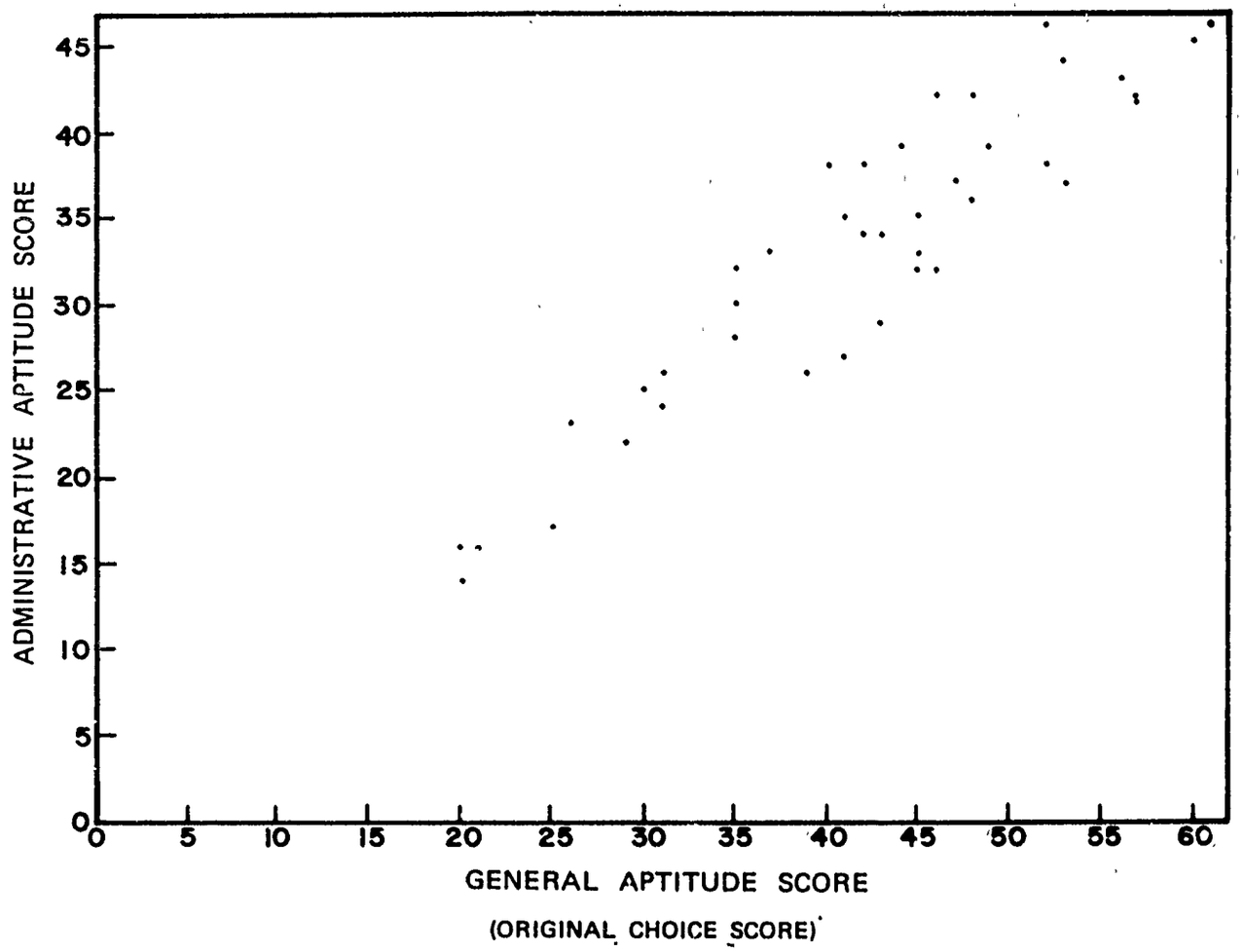


Figure 2a. Relation between General and Administrative Aptitude Scores based on number of correct answers during original administration of AQE--66 to 40 airmen.

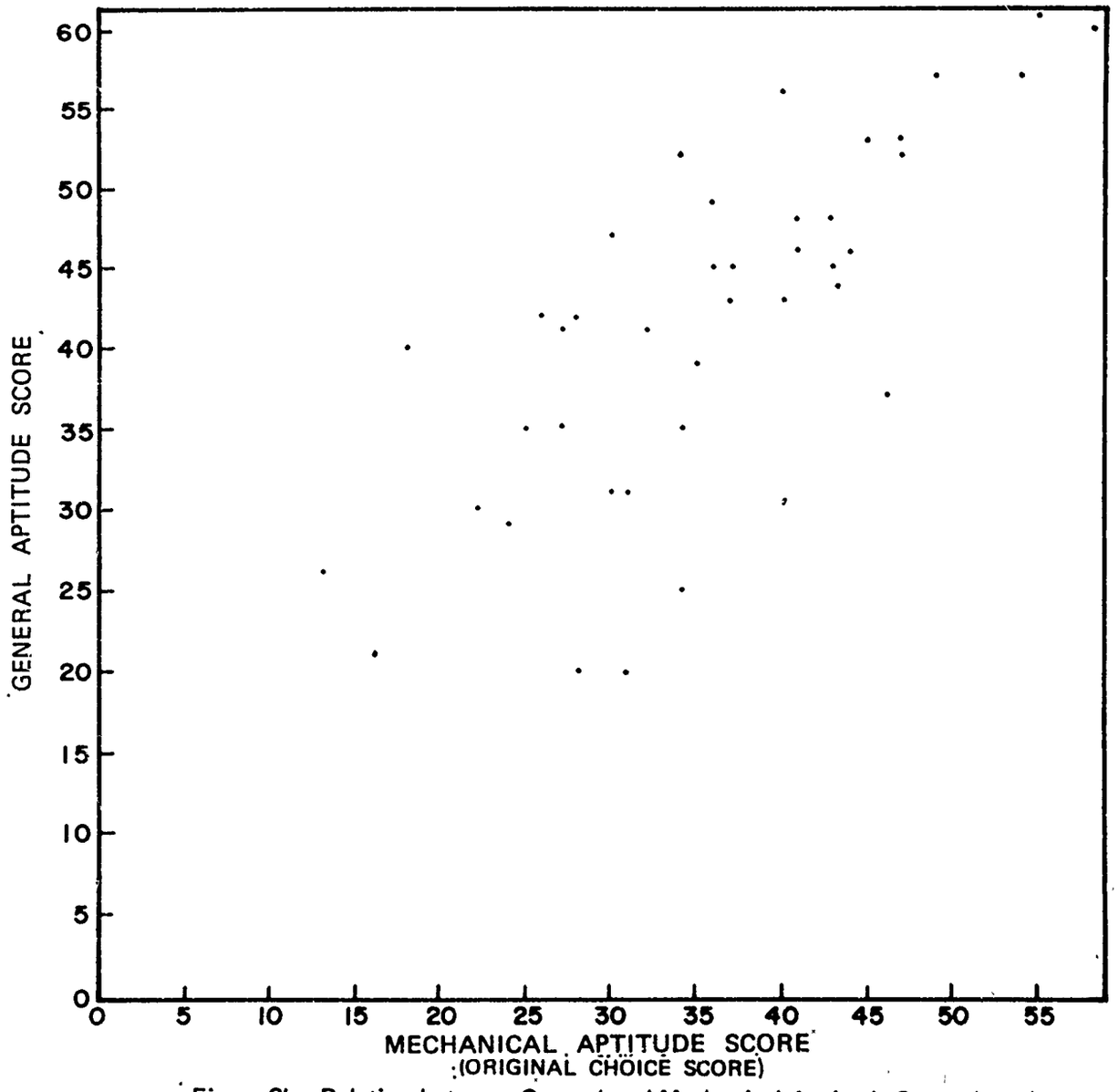


Figure 2b. Relation between General and Mechanical Aptitude Scores based on number of correct answers during original administration of AQE–66 to 40 airmen.

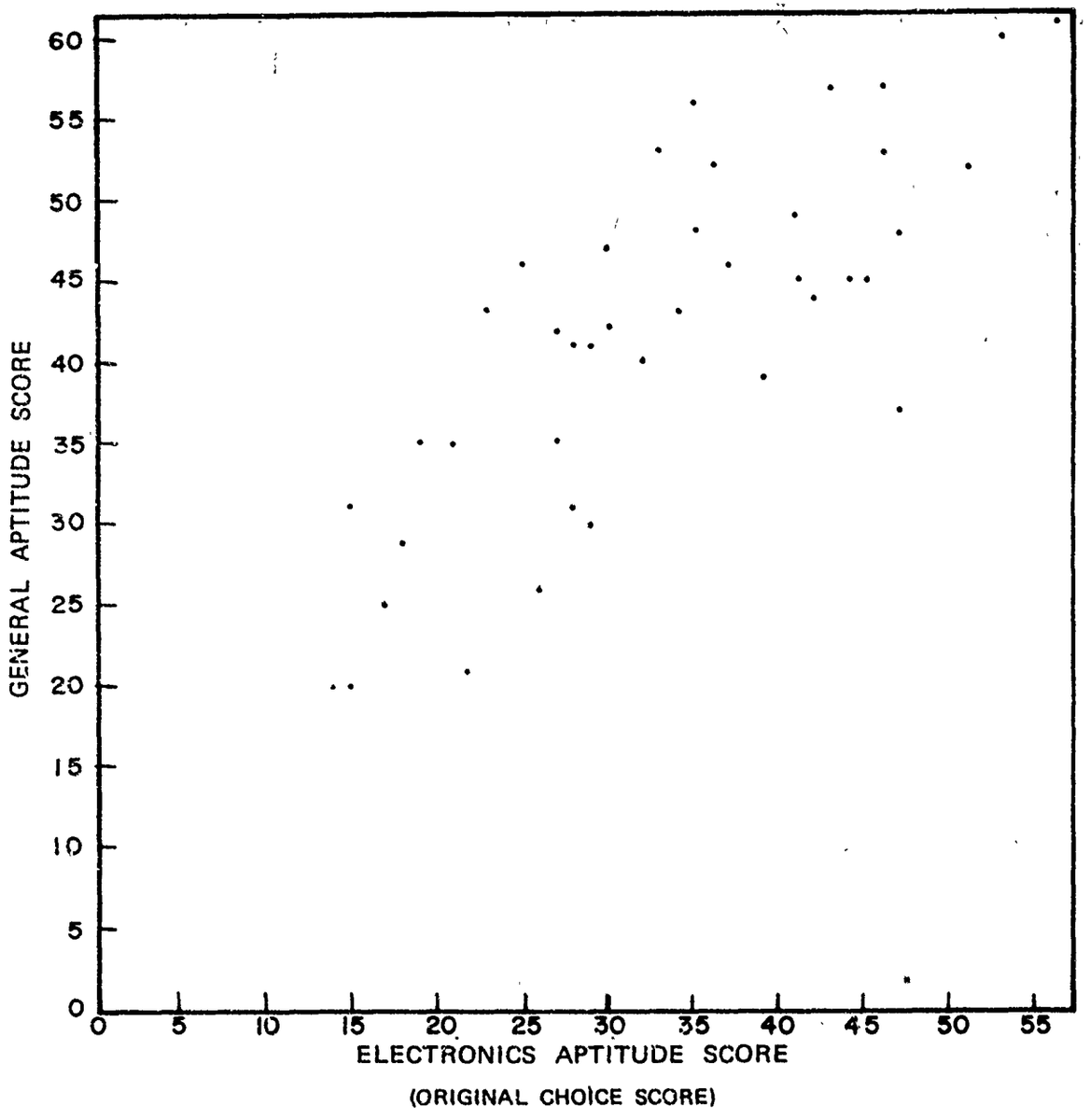


Figure 2c. Relation between General and Electronics Aptitude Scores based on number of correct answers during original administration of AQE–66 to 40 airmen.

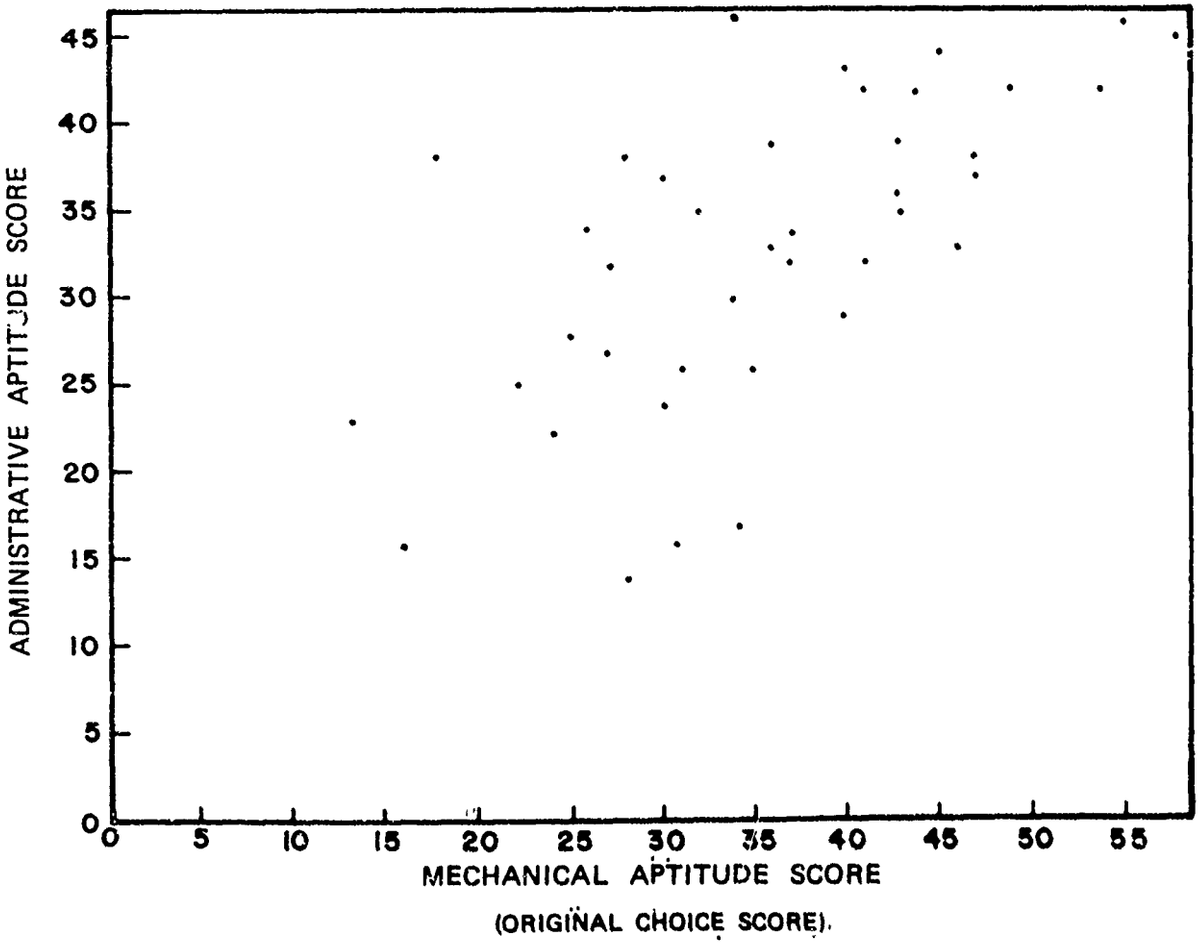


Figure 2d. Relation between Administrative and Mechanical Aptitude Scores based on number of correct answers during original administration of AQE–66 to 40 airmen.

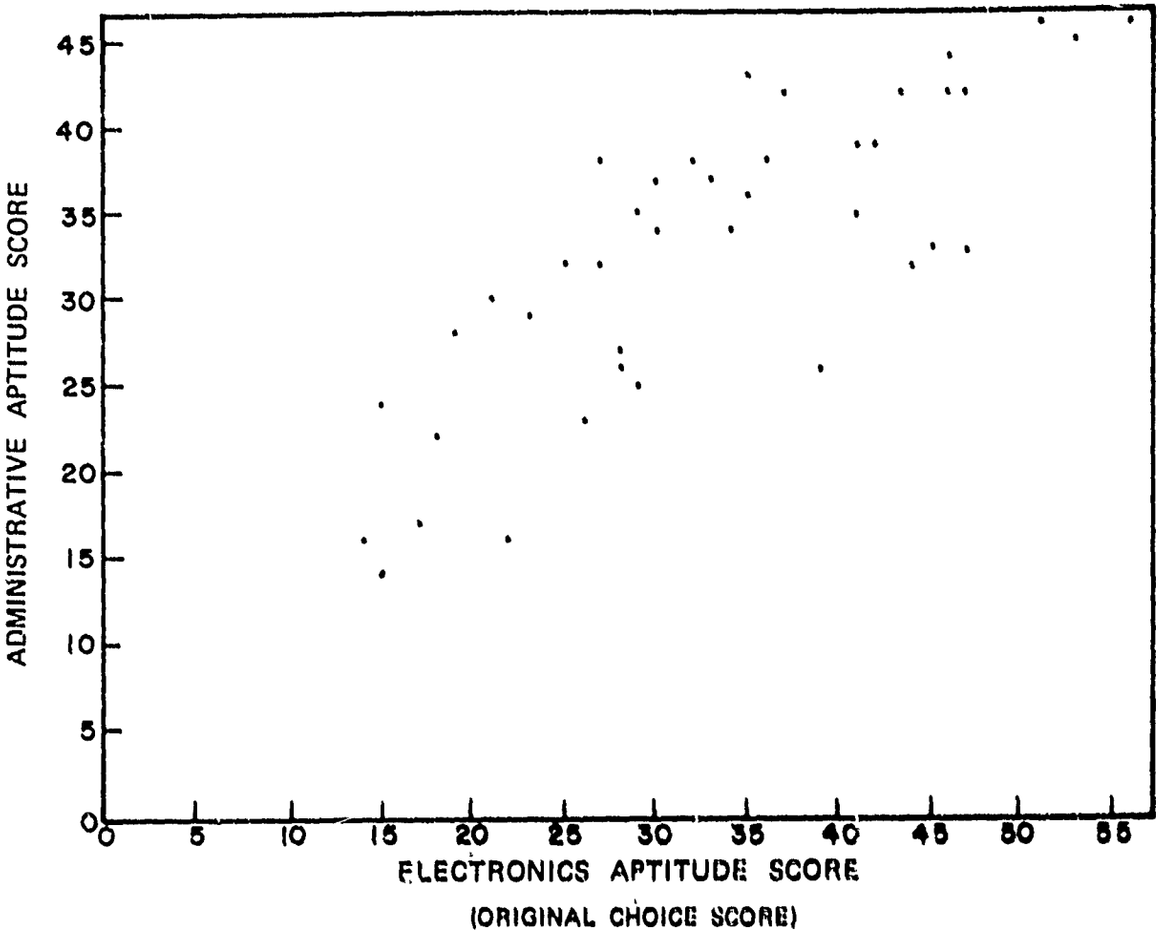


Figure 2e. Relation between Administrative and Electronics Aptitude Scores based on number of correct answers during original administration of AQE–66 to 40 airmen.

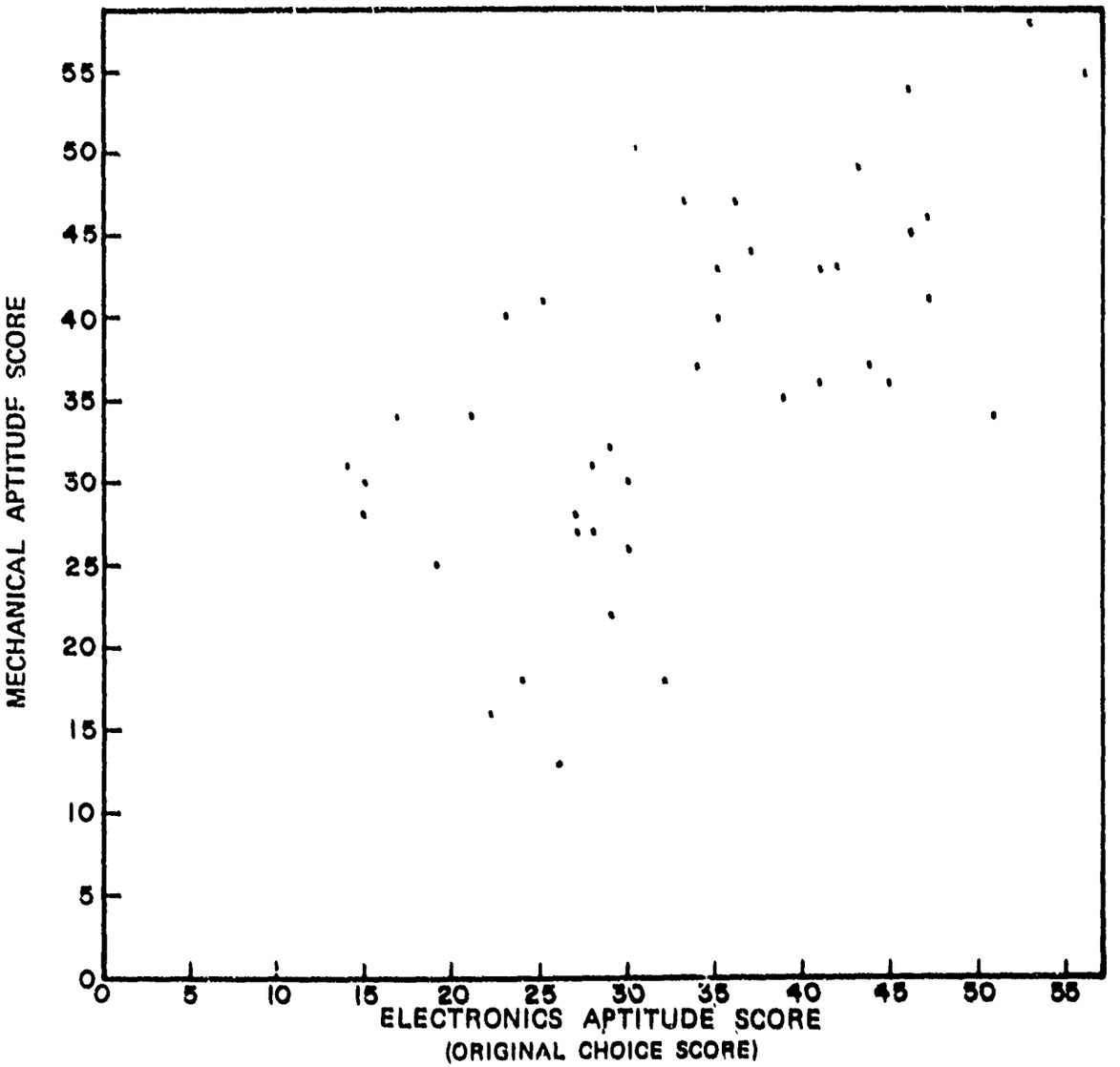


Figure 2f. Relation between Mechanical and Electronics Aptitude Scores based on number of correct answers during original administration of AQE–66 to 40 airmen.

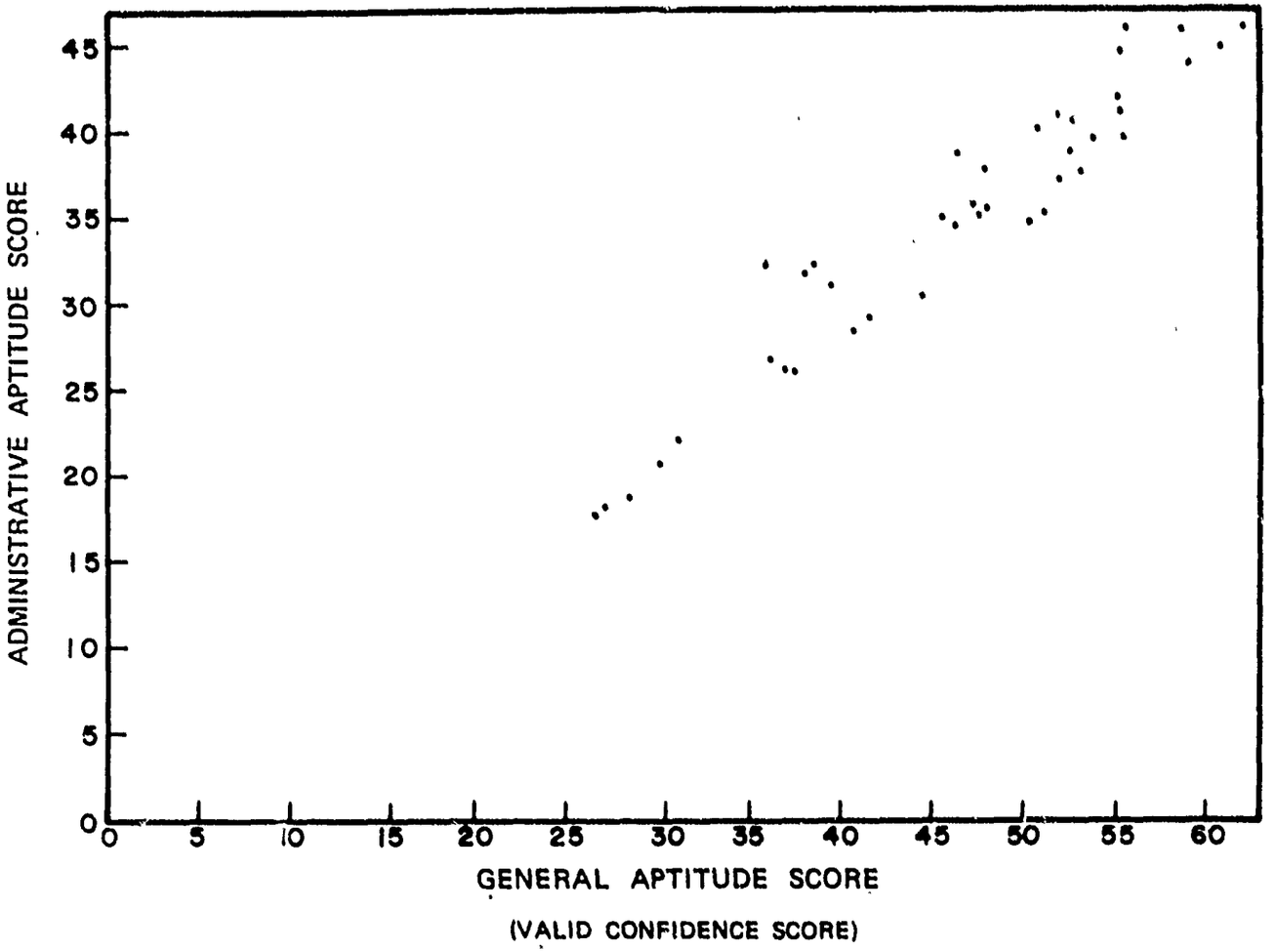


Figure 3a. Relation between General and Administrative Aptitude Scores based on Valid Confidence Scores from experimental administration of AQE–66 to 40 airmen.

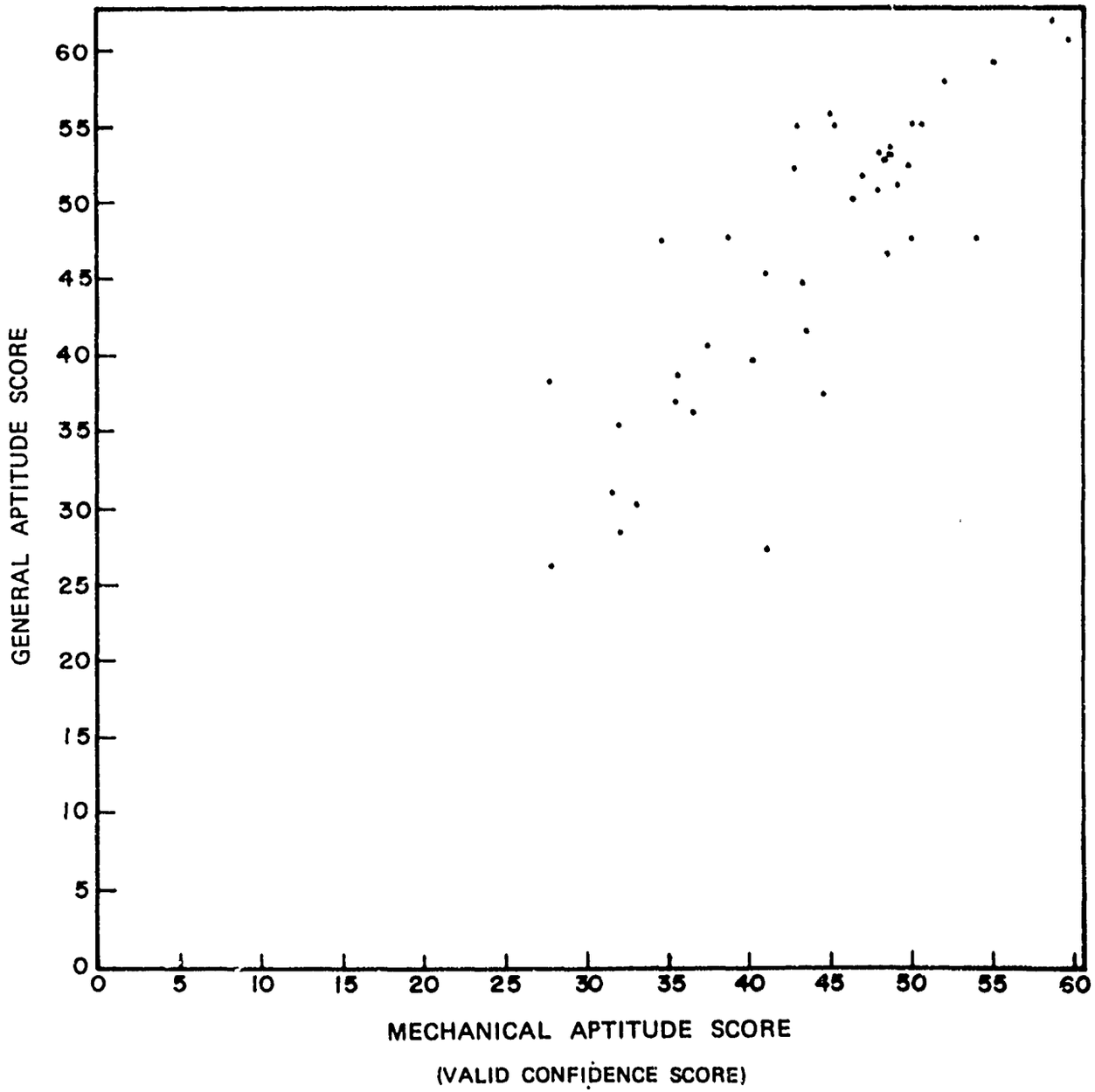


Figure 3b. Relation between General and Mechanical Aptitude Scores based on Valid Confidence Scores from experimental administration of AQE--66 to 40 airmen.

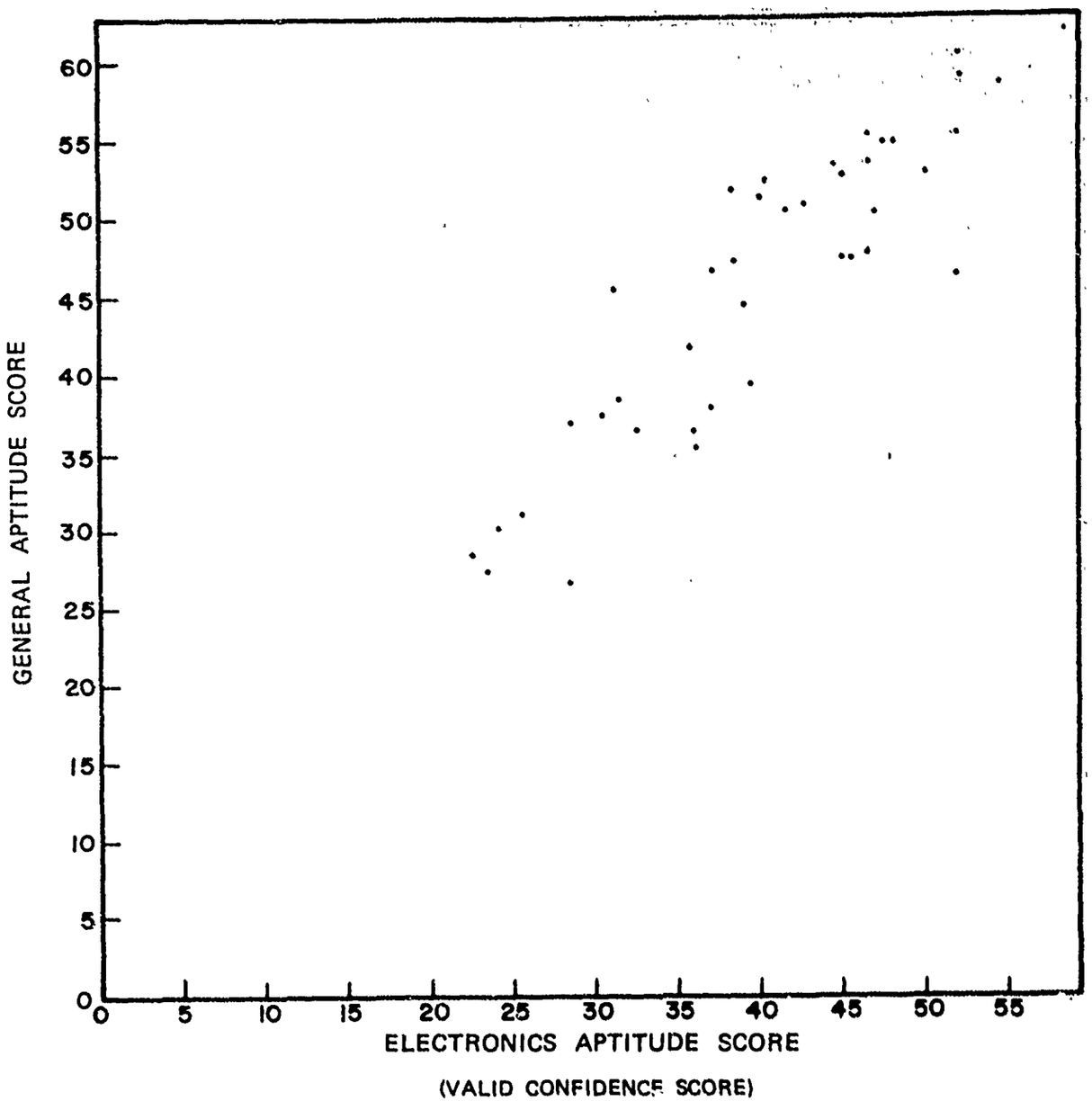


Figure 3c. Relation between General and Electronics Aptitude Scores based on Valid Confidence Scores from experimental administration of AQE–66 to 40 airmen.

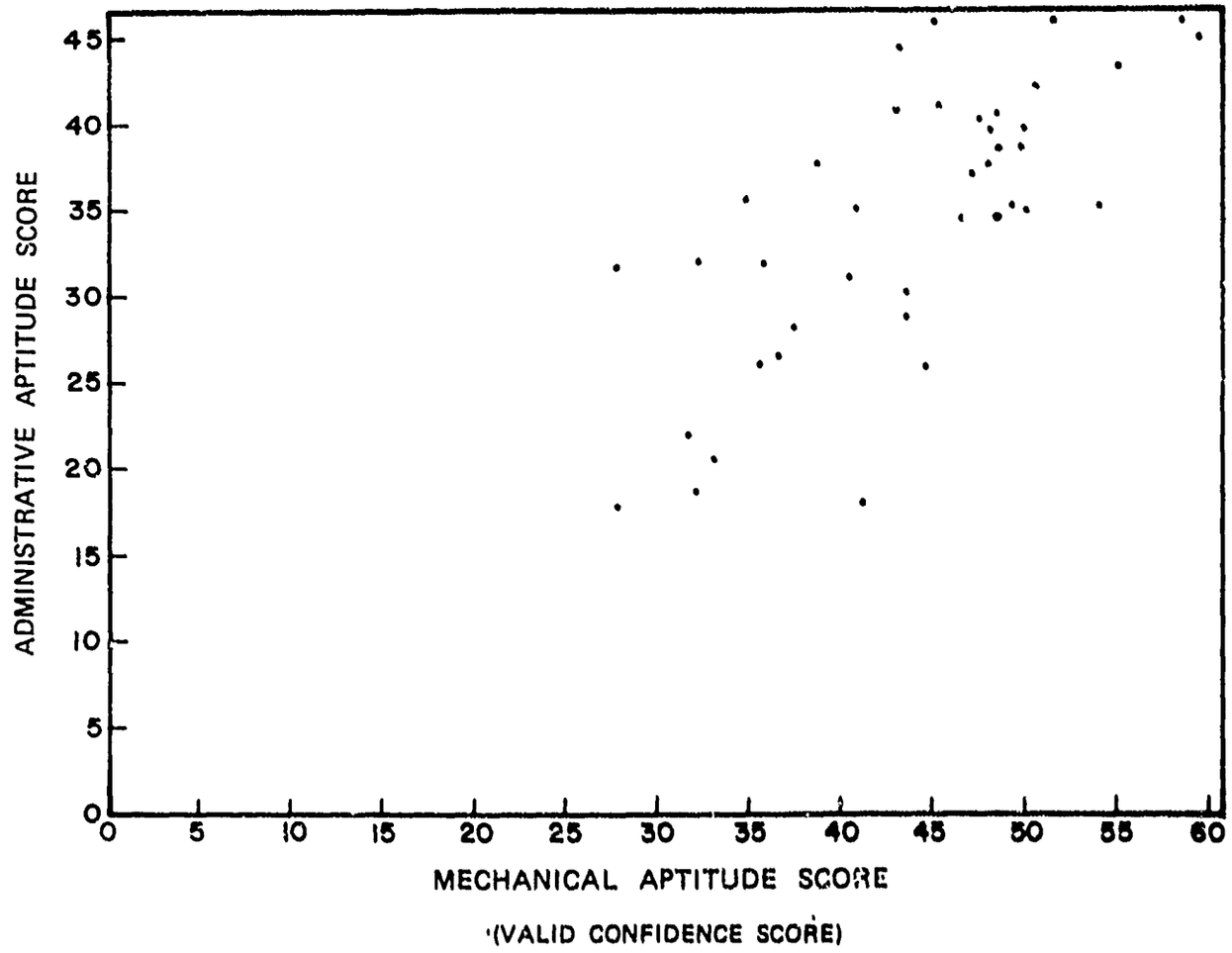


Figure 3d. Relation between Administrative and Mechanical Aptitude Scores based on Valid Confidence Scores from experimental administration of AQE–66 to 40 airmen.

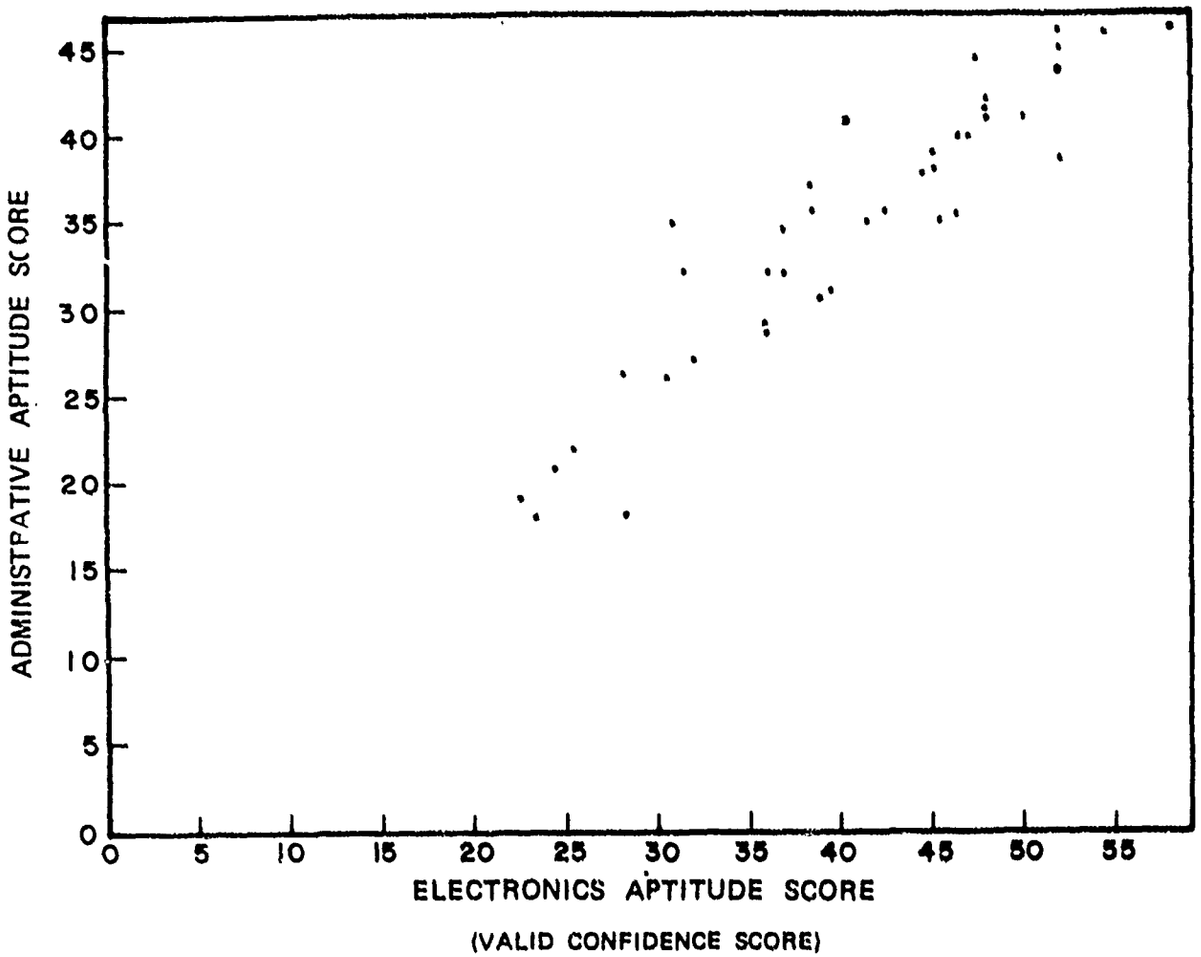


Figure 3e. Relation between Administrative and Electronics Scores based on Valid Confidence Scores from experimental administration of AQE--66 to 40 airmen.

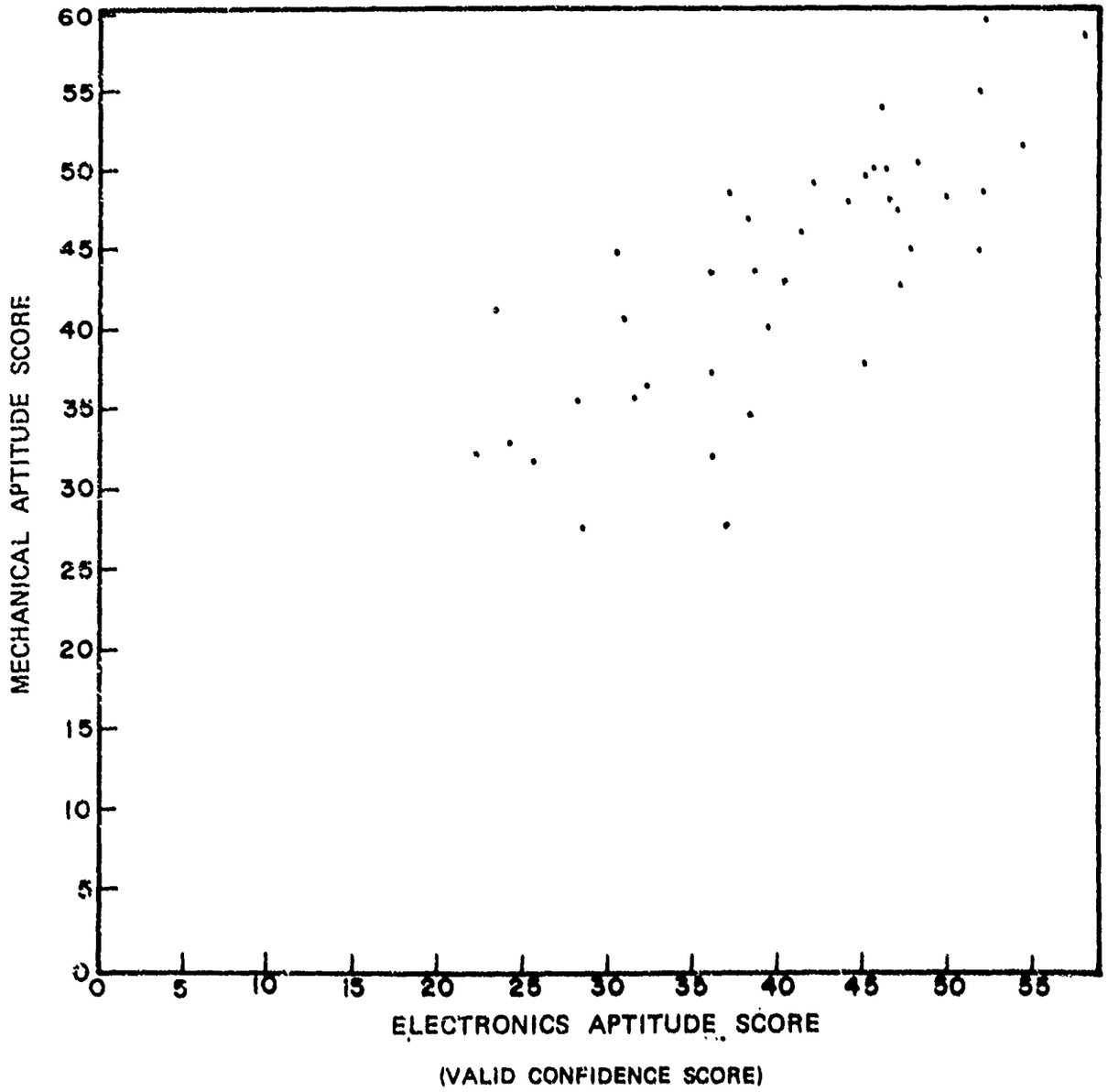
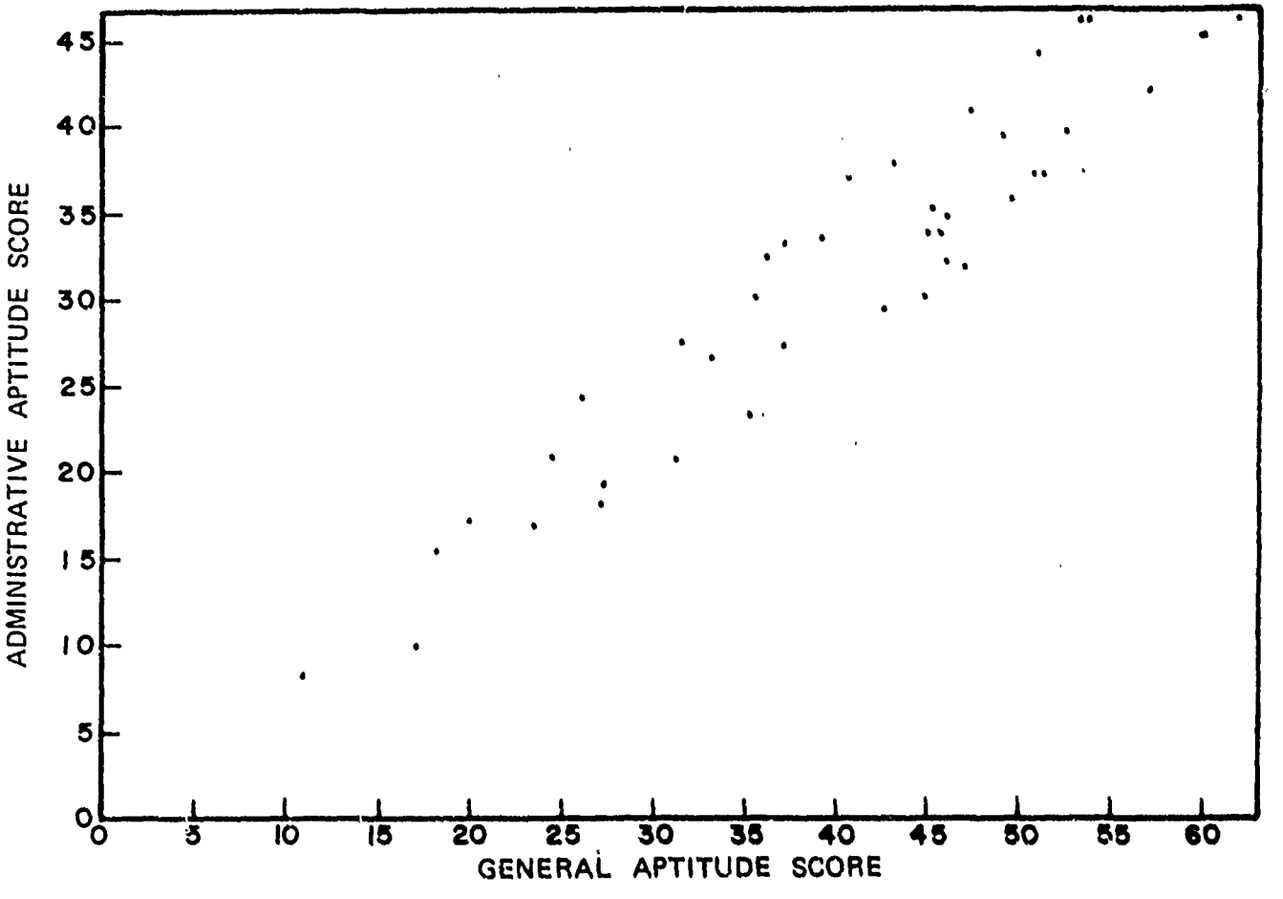


Figure 3f. Relation between Mechanical and Electronics Scores based on Valid Confidence Scores from experimental administration of AQE–66 to 40 airmen.



(TOTAL AMOUNT OF CONFIDENCE ON CORRECT ANSWERS)

Figure 4a. Relation between General and Administrative Scores based on total amount of confidence placed on correct answers during experimental administration of AQE–66 to 40 airmen.

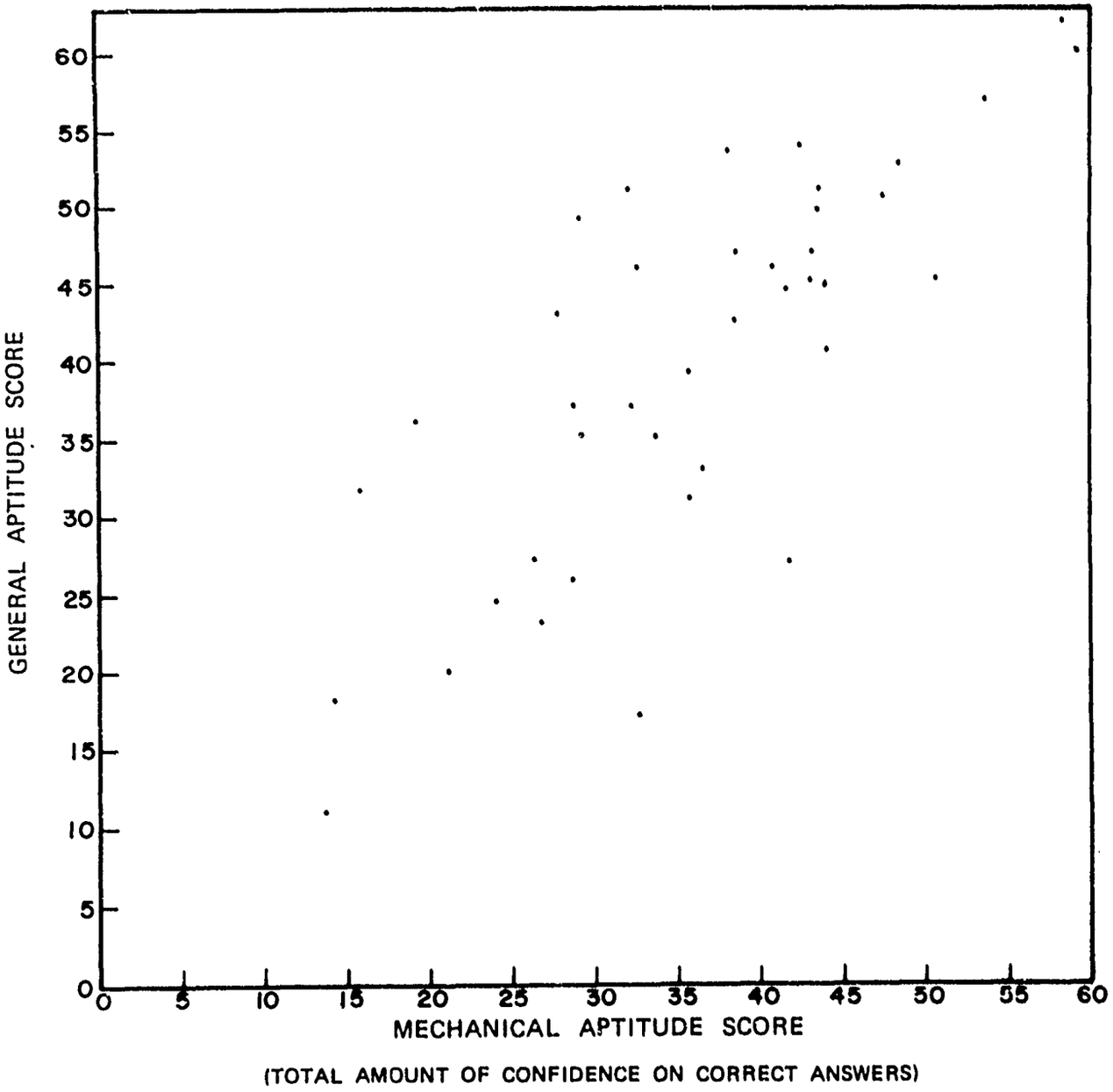


Figure 4b. Relation between General and Mechanical Scores based on total amount of confidence placed on correct answers during experimental administration of AQE–66 to 40 airmen.

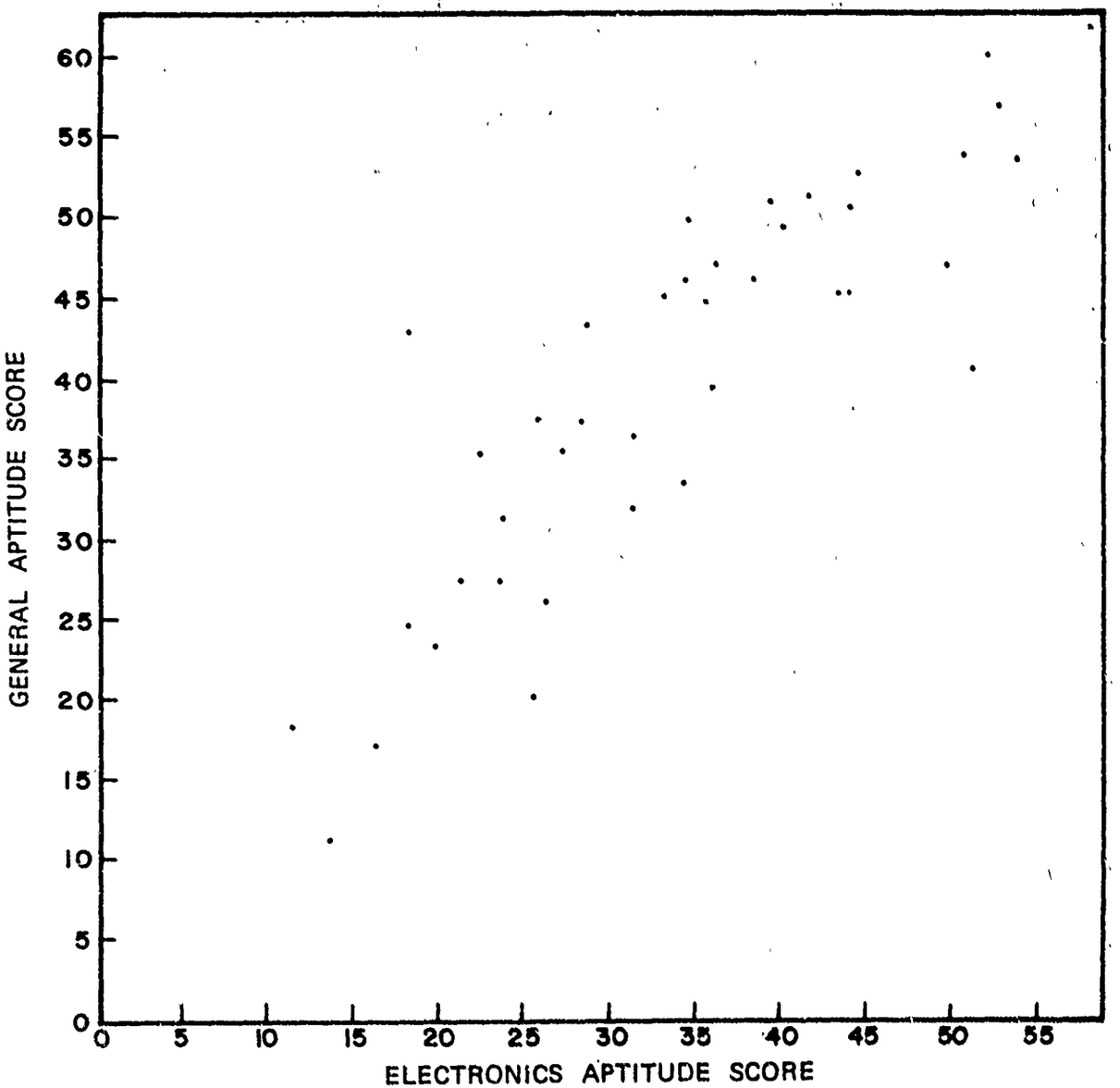


Figure 4c. Relation between General and Electronics Scores based on total amount of confidance placed on correct answers during experimental administration of AQE–66 to 40 airmen.

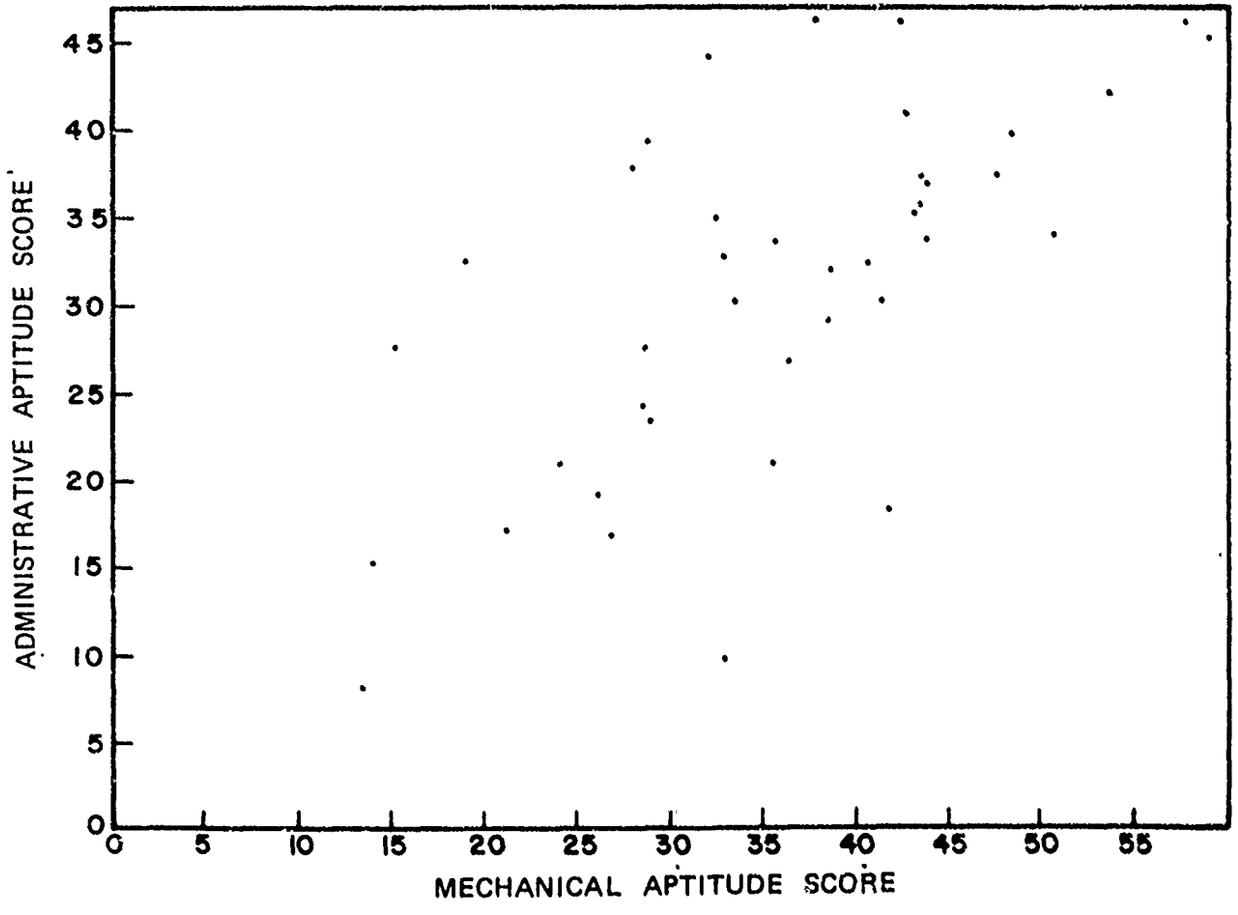


Figure 4d. Relation between Administrative and Mechanical Scores based on total amount of confidence placed on correct answers during experimental administration of AQE–66 to 40 airmen.

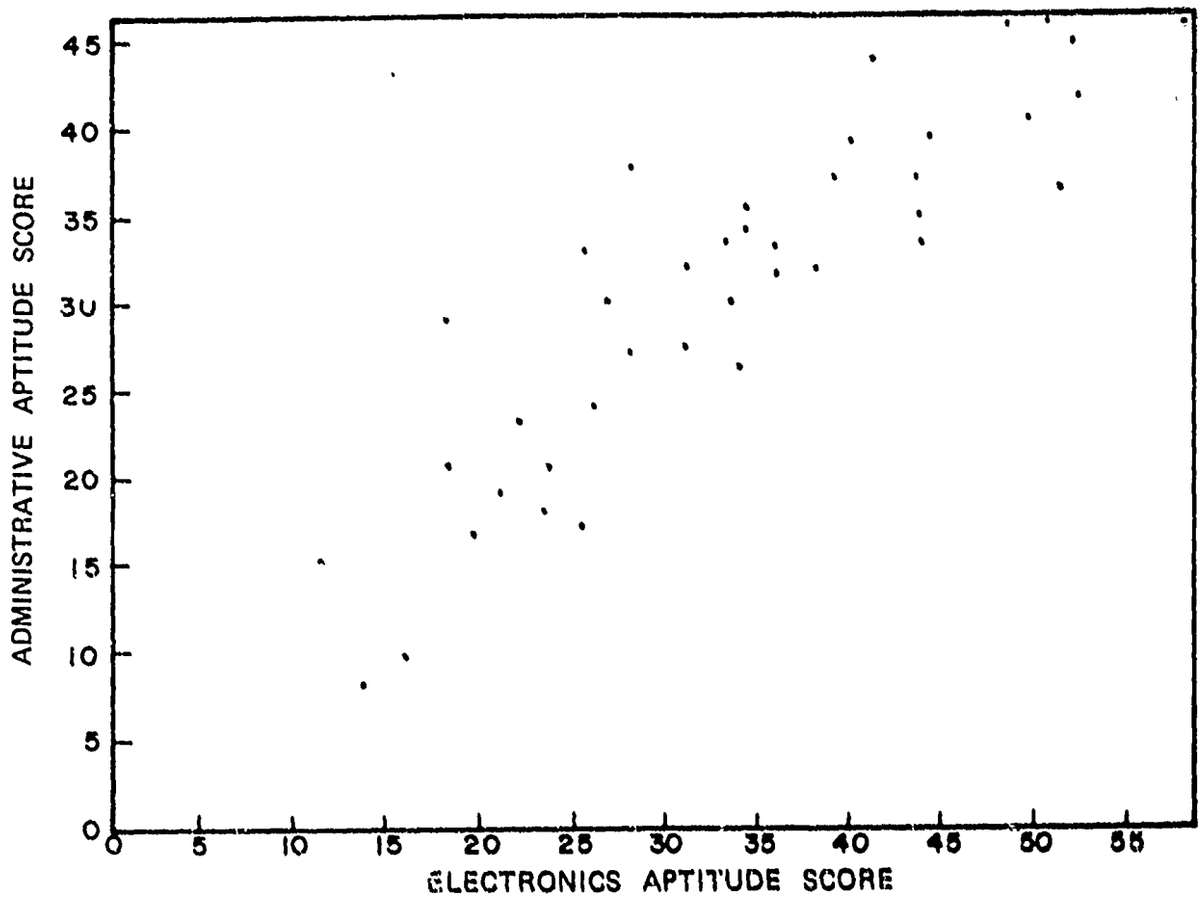


Figure 4e. Relation between Administrative and Electronics Scores based on total amount of confidence placed on correct answers during experimental administration of AQE–66 to 40 airmen.

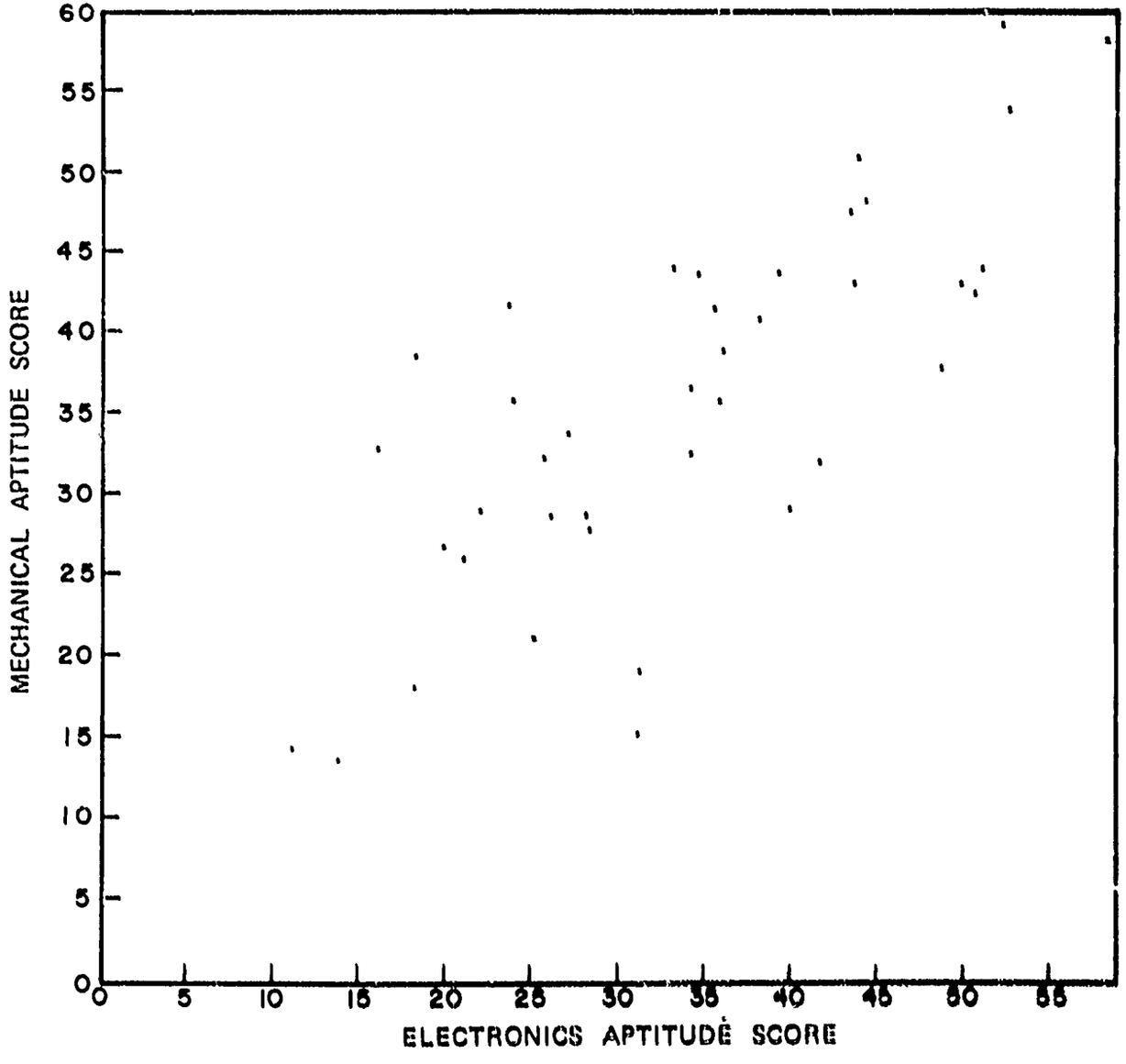


Figure 4f. Relation between Mechanical and Electronics Scores based on total amount of confidence placed on correct answers during experimental administration of AQE–66 to 40 airmen.

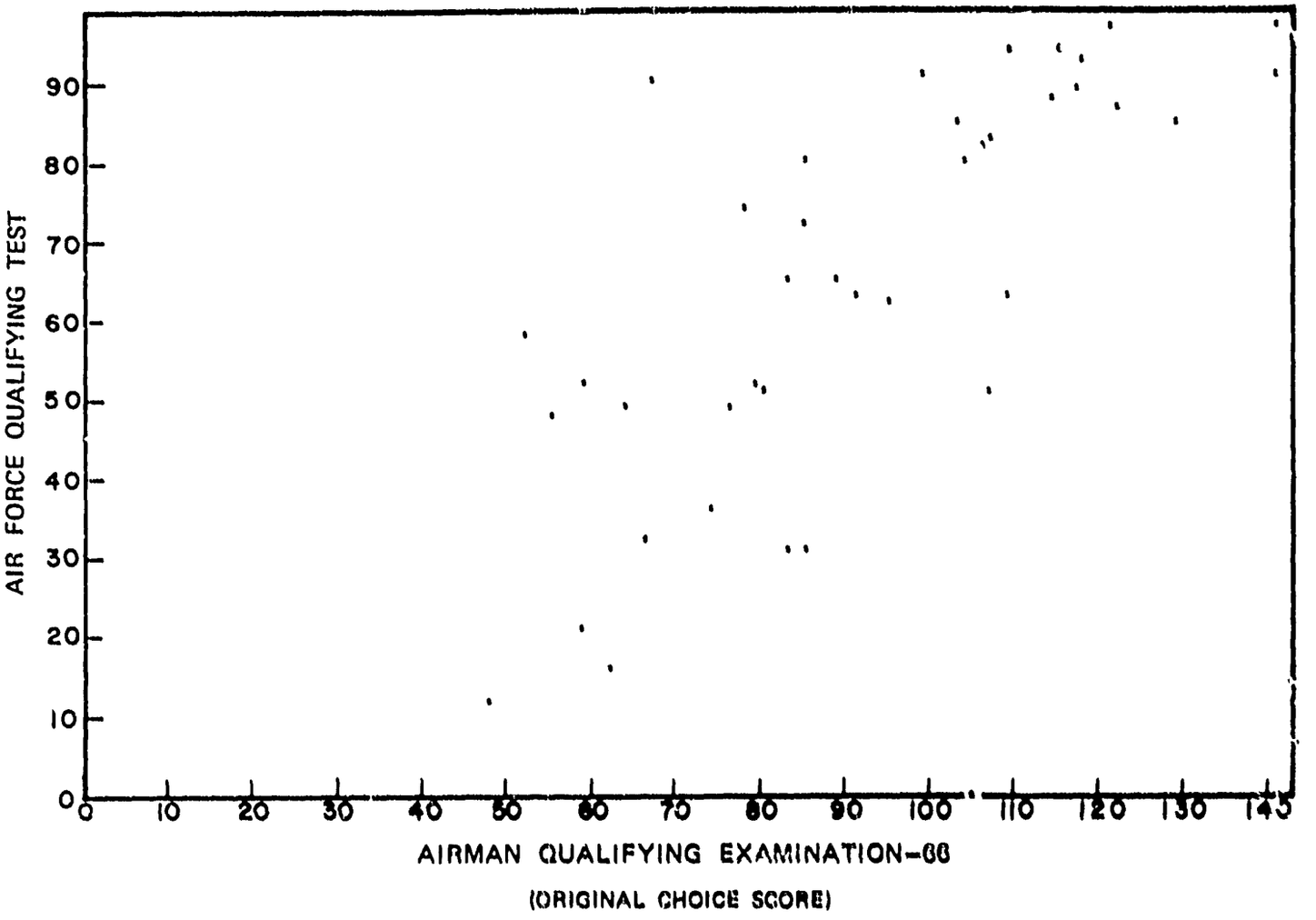


Figure 5. Relation between Air Force Qualifying Test Score and total number of correct answers to 150 multiple-choice items during original administration of AQE-66 to 40 airmen.

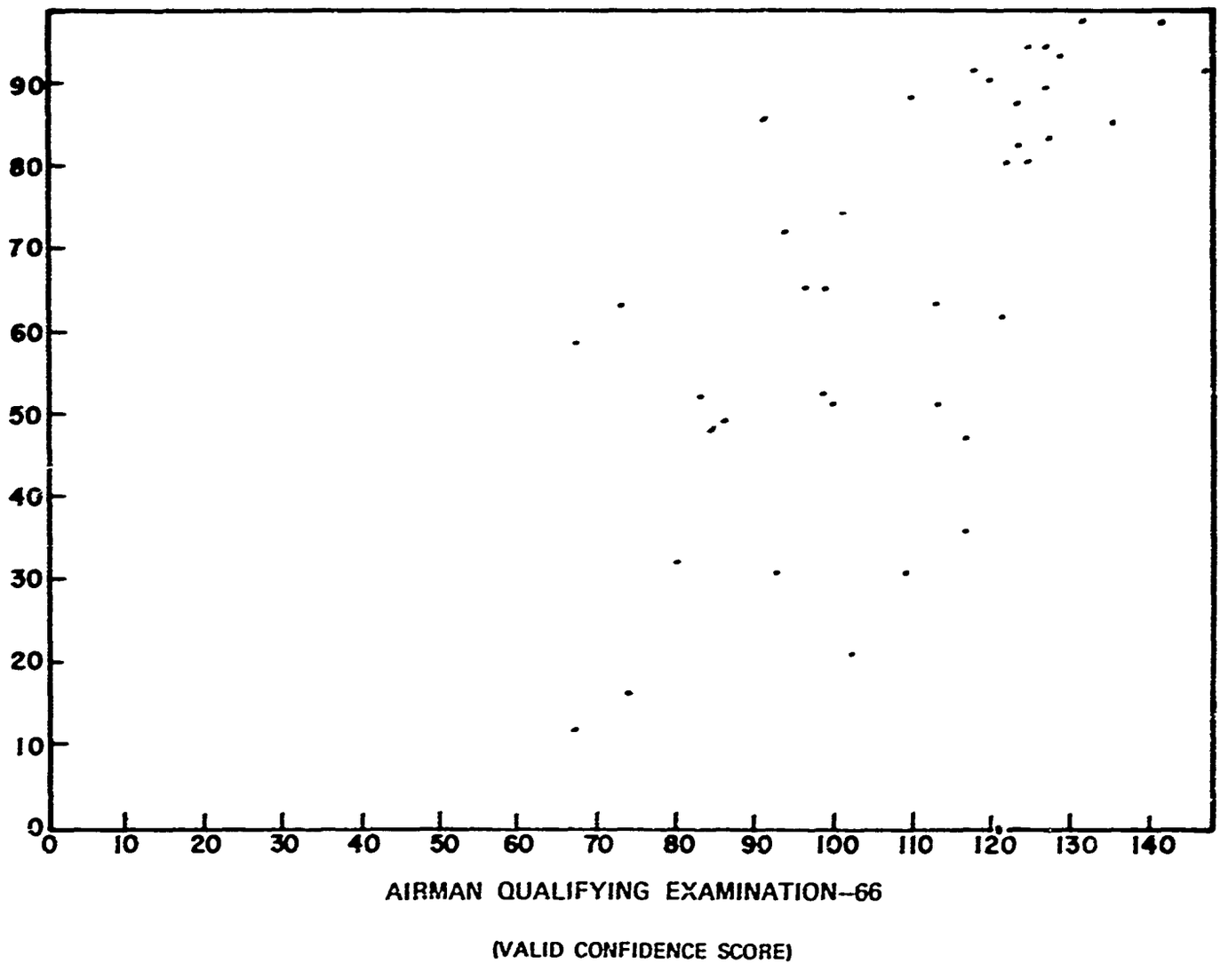


Figure 6. Relation between Air Force Qualifying Test Score and Valid Confidence Score for 150 multiple-choice items from experimental administration of AQE–66 to 40 airmen.

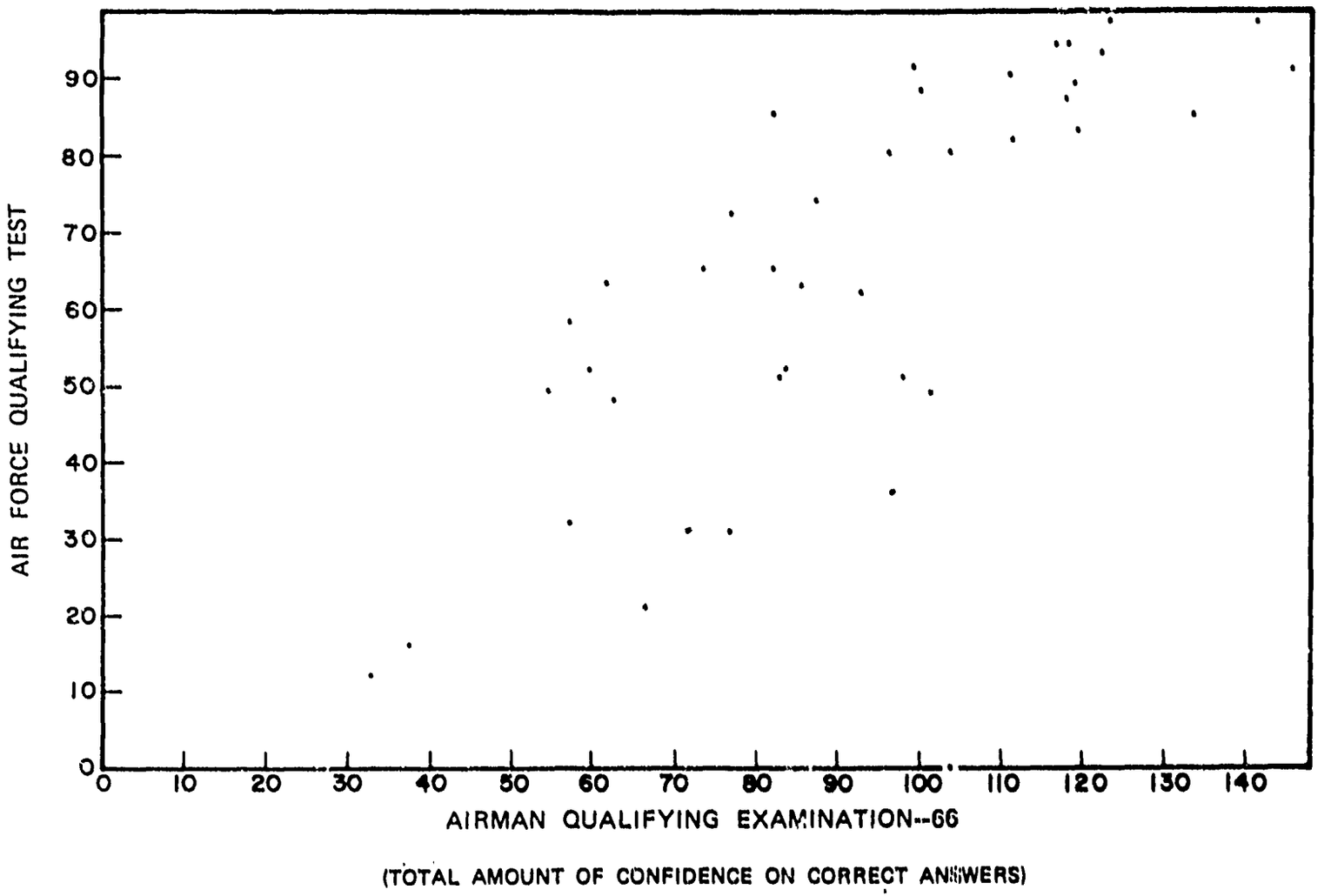


Figure 7. Relation between Air Force Qualifying Test Score and total amount of confidence placed on correct answers to 150 multiple-choice items during experimental administration of AQE-66 to 40 airmen.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) The Shuford-Massengill Corporation One Wallis Court Lexington, Massachusetts 02173	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

AIRMAN QUALIFYING EXAMINATION-66 ADMINISTERED AS A CONFIDENCE TEST

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific INTERIM

5. AUTHOR(S) (First name, middle initial, last name)

Emir H. Shuford, Jr. & H. Edward Massengill, Jr.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1968	13	14

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AF 49(638)-1744 b. PROJECT 9720F-9719 6154501R c. 681313	SMC R-12 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFOSR 68-2162

10. AVAILABILITY/LIMITATION NOTICES

1. This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
TECH, OTHER	Air Force Office of Scientific Research 1400 Wilson Boulevard (SRLB) Arlington, Virginia 22209

13. ABSTRACT

Airman Qualifying Examination-66 was readministered as a Valid Confidence test to 61 basic airmen.

Airmen understood the method of confidence testing and yielded data containing information over and above that available from choice testing. There is no evidence that confidence testing disrupted the test-taking process.

Wide individual differences were observed in airmen's ability to evaluate information. Observed patterns of confidence indicate that airmen would be guessing on about one-fourth of the items if AQE-66 had been administered as a choice test.

Confidence test administration served to increase reliability of AQE-66 to the extent that it was equivalent to a choice test about twice as long as the current test.

Confidence test administration served to increase predictive validity of AQE-66 as measured by the correlation between AQE-66 and AFQT.

DD FORM 1473

UNCLASSIFIED

Personnel Selection and Classification
Guessing
Decision-theoretic psychometrics
Valid Confidence Testing
test reliability
test validity
Air Force Qualifying Test